U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-288 320

Human Health Damages from Mobile Source Air Pollution: Additional Delphi Data Analysis. Volume II

California State Air Resources Board, Sacramento

Prepared for

Corvallis Environmental Research Lab, OR

Jul 78

United States
Environmental Protection
Agency

Environmental Research
Laboratory
Corvallis OR 97330

EPA-600/5-78-016b
July 1978

Research and Development

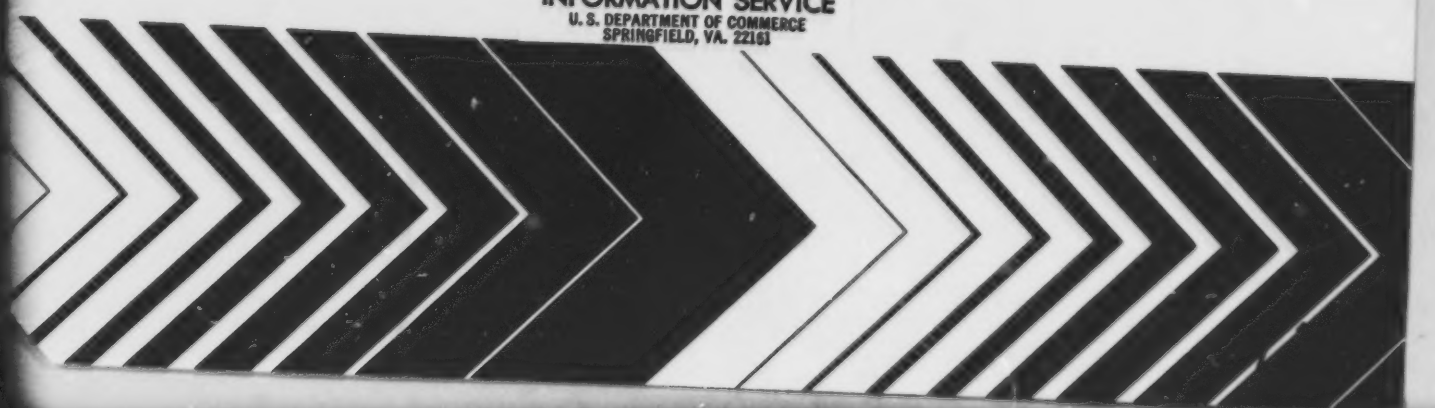
PB 288320

EPA

Human Health Damages From Mobile Source Air Pollution

Additional Delphi Data Analysis Volume II

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161



EPA-600/5-78-016b
July 1978

HUMAN HEALTH DAMAGES FROM MOBILE SOURCE AIR POLLUTION:

ADDITIONAL DELPHI DATA ANALYSIS - VOLUME II

by

Steve Leung
Eureka Laboratories, Inc.
401 N. 16th Street
Sacramento, California 95814

Norman Dalkey
University of California
Los Angeles, California 90025

Contract No. 68-01-1889

Project Officer

John Jaksch
Criteria and Assessment Branch
Corvallis Environmental Research Laboratory
Corvallis, Oregon 97330

CORVALLIS ENVIRONMENTAL RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
CORVALLIS, OREGON 97330

DISCLAIMER

This report has been reviewed by the Corvallis Environmental Research Laboratory, Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency and the California Air Resources Board, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/5-78-016b	2.
3. RECIPIENT'S ACCESSION NO. PB288320	
4. TITLE AND SUBTITLE: Human Health Damages from Mobile Source Air Pollution: Additional Delphi Data Analysis - Volume II	5. REPORT DATE: July 1978
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S): Steve Leung, Eureka Lab., Inc. 401 N. 16th St. Sacramento, CA 95814; Norman Dalkey, Univ. of Calif. L.A., CA 90024	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Contractor: California Air Resources Board, 1709 11th Street, Sacramento, CA 95814; Subcontractor: Eureka Laboratories, Inc. 401 N. 16th St., Sacramento, CA 95814	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO. EPA Contract No. 68-01-1889
12. SPONSORING AGENCY NAME AND ADDRESS: Corvallis Environmental Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Corvallis, Oregon 97330	13. TYPE OF REPORT AND PERIOD COVERED: Final Report
	14. SPONSORING AGENCY CODE: EPA/600/2

15. SUPPLEMENTARY NOTES

Volume I of this report is EPA-600/5-78-016a.

16. ABSTRACT

This report contains the results of additional analyses of the data generated by a panel of medical experts for a study of Human Health Damages from Mobile Source Air Pollution (hereafter referred to as HHD) conducted by the California Air Resources Board in 1973-75 for the U.S. Environmental Protection Agency (Contract No. 68-01-1889, Phase I).

The analysis focused on two topics: (1) assessment of the accuracy of group estimates and (2) generation of a model of the group estimate as a function of percent of population affected and degree of impairment.

Investigation of the first topic required a more thorough formulation of the statistical theory of errors as applied to group judgment than has been available up to now. This formulation is presented in Section 5 of the report. A major new feature of this theory is the postulation of a psychonumeric scaling on estimated numbers analogous to the psychophysical scaling of sensory magnitudes.

The investigation of the second topic and the application of the theory of errors to the data from the HHD studies are presented in Section 7.

This report was submitted by the California Air Resources Board in the fulfillment of Contract No. 68-01-1889 under the sponsorship of the Environmental Protection Agency. Work was completed as of September 30, 1976.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollutions Health Effects Delphi Study Dose-Response Decision Theory		

18. DISTRIBUTION STATEMENT: Unlimited	19. SECURITY CLASS *(This Report)*: Unclassified	21. NO. OF PAGES: 92
	20. SECURITY CLASS *(This page)*: Unclassified	22. PRICE: NTIS PAGE

EPA Form 2220-1 (Rev. 4-77)

U.S. GOVERNMENT PRINTING OFFICE: 1978—797-660/236 REGION 10

FOREWORD

Effective regulatory and enforcement actions by the Environmental Protection Agency would be virtually impossible without sound scientific data on pollutants and their impact on environmental stability and human health. Responsibility for building this data base has been assigned to EPA's Office of Research and Development and its 15 major field installations, one of which is the Corvallis Environmental Research Laboratory (CERL).

The primary mission of the Corvallis Laboratory is research on the effects of environmental pollutants on terrestrial, freshwater, and marine ecosystems; the behavior, effects and control of pollutants in lake systems; and the development of predictive models on the movement of pollutants in the biosphere.

A. F. Bartsch
Director, CERL

ABSTRACT

This report contains the results of additional analyses of the data generated by a panel of medical experts for a study of Human Health Damages from Mobile Source Air Pollution (hereafter referred to as HHD) conducted by the California Air Resources Board in 1973-75 for the U.S. Environmental Protection Agency (Contract No. 68-01-1889, Phase 1).

The analysis focussed on two topics: (1) assessment of the accuracy of group estimates and (2) generation of a model of the group estimate as a function of percent of population affected and degree of impairment.

Investigation of the first topic required a more thorough formulation of statistical theory of errors as applied to group judgment than has been available up to now. This formulation is presented in Section 5 of the report. A major new feature of this theory is the postulation of a non-linear response with estimated numbers similar to the non-linear relationship observed by psychologist between physical magnitudes and subjective estimates.

The investigation of the second topic and the application of the theory of errors to the data from the HHD studies are presented in Section 7. The hypothesis for this model is that the level of impairment scales on the logarithm of the concentration. The fit of the model to the data is surprisingly good. This model can be used to simplify additional Delphi studies of other pollutants. Thus, with the model, it is necessary to obtain estimates from panelists only of the base case, and the remaining cases can be predicted by the model.

This report was submitted by the California Air Resources Board in the fulfillment of Contract No. 68-01-1889 under the sponsorship of the Environmental Protection Agency. Work was completed as of September 30, 1976.

CONTENTS

Foreword iii
Abstract iv
List of Figures vi
List of Tables viii
Abbreviations and Symbols ix
Acknowledgments xi

1. Executive Summary 1
2. Conclusions 8
3. Recommendation 10
4. Introduction 11
5. Conceptual Background 12
    A. Theory of Errors and Group Judgment 12
    B. Expected Error 18
    C. The Psychonumeric Hypothesis 21
    D. Self-Evaluation 32
    E. Estimated Confidence Ranges 34
6. Research Methods 35
7. Results 38
    A. Theory of Errors 38
    B. Lognormality 41
    C. Logarithmic Scaling 45
    D. Correlation of Indices of Uncertainty 53
    E. An Estimation Model 64
8. Discussion 74

References and Notes 77

FIGURES

Number		Page
1.	Illustrative Distribution of Individual Responses	13
2.	Illustration of Bias and Random Error	14
3.	Distribution of Initial Answers	17
4.	Invariance of E/σ	19
5.	Cumulative Frequencies of z_m on Probability Scale	22
6.	Density Distribution of e^{z_m}	23
7.	Relative Frequency of Digits Occurring as Second Digits in in Almanac Tables (3114 numbers)	26
8.	Distribution of First Digits, Subject Responses (5,037 Responses)	27
9.	Average Standard Deviation as a Function of Log True	29
10.	Average Error as a Function of Log True	30
11.	Group Self-Rating	33
12.	Frequency Distribution of z Scores for all Best Estimates	43
13.	Observed Standard Deviation S as a Function of the Log Geometric Mean m for the Normal Population	46
14.	Average Standard Deviation vs Percent Population	48
15.	Average Standard Deviation s vs Average Mean m for Oxidant	49
16.	Average Standard Deviation s vs Average Mean m for Nitrogen Dioxide	50
17.	Average Standard Deviation s vs Average Mean m for Carbon Monoxide	51
18.	Illustration of Reduction in Correlation with Pooled Populations	56

Number		Page
19.	Average Estimated Interval vs Observed Standard Deviation	60
20.	Average Estimated Confidence Interval vs Average Log Mean	61
21.	Normalized Estimate $\hat{X}$ as a Function of the Parameter P	68
22.	Illustration of Cumulative Distribution	71
23.	Variation of $\bar{x}$ with Population	73

TABLES

Number		Page
1.	The Ratios of S^*/S and GM/Md for the Normal Population	40
2.	Average Error and Confidence Limits from Theory of Errors	41
3.	Average s and Average m of all Population Groups	52
4.	Correlation Between Self-Rating and Confidence Range	54
5.	Correlation Between R and Δy, for NO_2, Disability, Children	57
6.	Correlations Between $(\Delta\tilde{y},s)$, $(\Delta\tilde{y},\tilde{R})$, $(s,\tilde{R})$	58
7.	Average $\tilde{R}$ by Percent Population, Degree of Impairment and Pollutant Type	62
8.	Overall Averages for Δy, R and s	63
9.	$\Delta y/3.28s$ for Three Pollutants	63
10.	Normalized Estimates and Standard Deviation by Percent Population, Degree of Impairment and Pollutant Type	67

ABBREVIATIONS AND SYMBOLS

X_i -- individual response

x_i -- log X_i

i -- individual members of group ranging from 1 to n

T -- true answer

B_i -- a bias term

b_i -- log B_i

M_i -- mean of individual's distribution of responses

m_i -- mean of distribution of logarithm of individual's responses

μ -- theoretical mean of a distribution of log quantities

σ -- theoretical standard deviation

S -- observed standard deviation

s -- observed standard deviation of the logs

n -- total number of respondents or sample size

$D_i(x)$ -- density distribution of individual's responses

Ex -- expectation

R_i -- random error of the individual's response

r_i -- random error of the logarithmic of the individual's responses

z_m -- normalized form of the mean of the logarithmic of responses

Ψ -- psychological magnitude

HHD -- Human Health Damage

S^* -- best estimator of standard deviation

OX	-- oxidant
CO	-- carbon monoxide
NO_2	-- nitrogen dioxide
AE	-- theoretical average error
AE'	-- empirical average error
CL	-- theoretical confidence limits
CL'	-- empirical confidence limits
x^2	-- chi square
q_i	-- observed frequency in cell i
p_i	-- expected frequency in i
$\bar{Y}_u$	-- mean of upper limit
$\bar{Y}_l$	-- mean of lower limit
$\Delta\bar{y}$	-- difference between means of upper and lower limits, $\bar{Y}_u - \bar{Y}_l$
$\bar{R}$	-- average self-rating
r_{xy}	-- correlation for pooled population
$C\bar{V}(x,y)$	-- a generalized covariance of the means of subpopulation
F	-- fraction of the population affected
c	-- dosage
$\phi(z)$	-- normal density function
z	-- normalized variate
Dc	-- Discomfort
Da	-- Disability
I	-- Incapacity

ACKNOWLEDGEMENTS

This study was an extension to Contract No. 68-01-1889 from the U.S. Environmental Protection Agency. Assistance in planning program objectives and program direction was given by John A. Jaksch of the Environmental Protection Agency's Corvallis Environmental Research Laboratory, Criteria and Assessment Branch.

The authors are indebted to the staff members of the Air Resources Board. Appreciation is extended to Jack Suder, Laura Storey Dick and Kingsley Macomber for their assistance in administrative and contractual matters.

Grateful acknowledgement is made to John R. Kinosian, Chief of the Technical Services Division, and John Holmes, Chief of the Research Division for valuable advice and encouragement.

SECTION 1

EXECUTIVE SUMMARY

This report contains additional analyses of data obtained in a Delphi study of dose-response relationships for air pollutants conducted by the California Air Resources Board for the Environmental Protection Agency*. The study involved collection of estimates from a panel of 14 medical experts of the dosage (concentration of a given pollutant experienced for one hour) required to produce a given level of impairment in a given fraction of a specific population at risk. The panel judgments generated in the study constitute one of the best and most complete sets of data that have been collected in a Delphi study of health effects. The additional analysis reported in the present supplement has two aims: (a) formulating a more complete and rigorous treatment of the reliability of the data, and (b) a more complete investigation of the underlying model determining the relationships between dose, type of disability, and fraction of population affected.

I. RELIABILITY

Data generated by Delphi studies are not equivalent to data generated by statistical sampling procedures. Panels are not selected for representativeness, but for expertise. Responses by panel members are taken to be their best judgments on a question, and should be scored for correctness, not for what they convey concerning the panelists. As a result, standard statistical analyses are not directly applicable to evaluating the excellence of Delphi results. Applicable standards are "quasi-statistical" in that indices are defined in statistical lenguage, but the criteria are derived from empirical studies.

To date the most extensive and relevant empirical studies have involved experiments with university upper-class and graduate students, with general

---

* Referred to in the following as the HHD (Human Health Damages) Study.

information and short-range prediction questions. The material was selected so that the students would have some information on which to base on estimate, but they were not expected to know the precise answers. The essential property of the questions is that the accuracy of the responses could be measured objectively.

The basic issue with regard to the applicability of the results of these experiments to the HHD study is the closeness of the analogy between the estimation process exhibited by the student subjects and the estimation process exhibited by the panel of medical experts. The best evidence would consist of assessing the panel estimates against extensive field studies with actual pollution situations. Lacking this field data, several less definitive considerations can be examined with the present data.

To provide a firm mathematical basis for the analyses, the standard theory of errors has been extended to include the case of log normal distributions, with the panel distribution of estimates assumed to arise from independent individual distributions. In addition to random error, provision is made for systematic error or bias. A general conclusion from the empirical studies is that bias is a larger contributor to the total error than random variability.

A. Log-normality of Distributions. The criteria for evaluating Delphi estimates that have been deduced from the experimental data are based on the assumption that the distributions of responses are log-normal. In the experiments, the distributions were log-normal to a high degree of approximation. The data from the HHD study are compatible with the assumption that the medical panel's estimates are log-normally distributed, providing that allowance is made for observed dependencies among individual panelist's estimates for closely related cases.

B. Observed and Estimated Standard Deviations and Medians. The critical statistic derived from the laboratory studies is the ratio of the sample standard deviation to the median. For a log-normal distribution, the median coincides with the geometric mean. The maximum likelihood estimate for the geometric mean is the anti-logarithm of the mean of the logarithms of the responses. The maximum likelihood estimate of the sample standard deviation is obtained by first computing the standard deviation of the logarithms of the estimates, and then computing the standard deviation of the original estimates from that of the logs.

Examination of the HHD data indicates that there is a large enough discrepancy between the statistic standard deviation/median computed from the raw data and the statistic computed from the log transform data to recommend that evaluations of panel estimates be based on the statistic computed from the log transform, not on the statistic derived from the raw data.

C. Conclusions from Pure Sampling Theory. Although the panel estimates cannot be considered as sampling data, some conclusions can be drawn treating the data as if it were the results of a sampling procedure. In particular, some upper bounds can be set on the reliability of the data.

Assuming that all the variance in the estimates arises from random variability, average expected errors for the three pollutants are 10% for Oxidant, 21% for Nitrogen Dioxide, and 18% for Carbon Monoxide, with confidence ranges of ± two standard deviations of ±28%, ±63%, and ±53% for the three pollutants respectively. These are gross averages intended to show the order of magnitude of expected errors based on random sampling theory alone. They are weak lower bounds to the expected error that would be derived using the empirical relationships between standard deviation and error.

D. Correlations Between Indices of Uncertainty. Three different indices of the panelists' confidence in their answers can be derived from the data. The panelists were asked to rate their confidence in each estimate on a scale from 1 to 10, where 1 meant "sheer guess" and 10 meant "I know the answer". In addition, each panelist estimated high and low bounds on their estimates where the high bound was defined as the concentration such that no more than 5% of cases would exceed the bound; and the low bound was defined as the concentration such that no more than 5% of cases would be lower. The self-rating and the high and low bounds are explicit judgments by the panelists of the reliability of their responses. In addition, the standard deviation is an implicit measure of the degree of certainty. In the experimental studies, the correlation between the average self-rating and the standard deviation was .67.

For the HHD data, correlations between these three indices were rather low, and in some cases were of the opposite sign from what would be expected. For example, it would be expected that the correlation between self ratings and estimated confidence ranges (the difference between the upper and lower bounds) would be negative since a wide range indicates low confidence. In some cases, the correlation between these two indices were positive.

Part of the explanation for the anomalous findings is that the panelists were estimating the upper and lower bounds as fractions of their original estimates, whereas in general, they were determining their original estimates as if the concentrations were scaled in a logarithmic manner.

The analyses, then, lends some support to the assumption that both the self-ratings and the upper and lower bounds are related to the degree of credence that can be placed in the estimates. However, the data does not support the assumption that the upper and lower bound estimates can be used to establish firm confidence limits for purposes of policy formulation.

E. Relation of Indices of Uncertainty to Problem Characteristics.

One question of interest is whether the panel showed systematic relationships between the indices of confidence and the major variables -- type of pollutant, level of impairment, percentage of population affected, and population type. One clear result is that the panel is less certain in their estimates for Nitrogen Dioxide than they are for Oxidant or Carbon Monoxide. This holds true for all three indices, self-rating, confidence bounds and standard deviation. There also appears to be an interesting positive relationship between self-ratings and percentage population. The panel expressed greatest confidence in their estimates for the 90% population, less in estimates for 50%, and least in estimates for 10%. This effect could indicate an interaction between the degree of confidence and the percentage estimate; percent population may represent another scale of "certainty".

The major effect observed with the confidence estimates (difference between upper and lower bounds) is an increase roughly proportional to the size of the concentration being estimated. Thus, to a first approximation, the panel is expressing a similar degree of uncertainty for the basic cases.

With regard to the relationship between self-rated confidence and populatio type, the clearest conclusion is that the panel felt most confident about their estimates concerning the normal population. For most of the other population types the average self-ratings were relatively uniform except for what appear to be special cases. For Oxidant, the special cases are asthma, chronic lung obstruction, and viral bronchitis. For the first two, the self ratings were high, which might indicate an affirmation that the cases were "serious" for oxidant effects. The average rating for viral bronchitis was distinctly low.

II. SUBSTANTIVE ANALYSES

The purpose of the substantive analysis was to test the hypothesis that the panelists were implicitly using a relatively simple model to formulate

their estimates for the various cases associated with a given pollutant and population type. Specifically, the hypothesis is that each panelist treats one of the cases (degree of impairment and percent population) as a "base" case and derives his estimates for the other cases by a systematic modification of the base case.

A beginning for such a model is furnished by the assumption that the panelists view the effects of air pollution as a threshold phenomenon, where within a given population there is a distribution of concentration levels for the onset of a given symptom. This submodel is rounded out be the assumption that the distribution of thresholds is normal on the logarithm of the concentration. This model was used to analyse the data in the original study, and found to fit the data rather well. However, the test was somewhat weak in that only three points were used to fit the cumulative dose-response curves for a given level of impairment.

In the present analysis, the hypothesis is extended to include both the dose-response curve and the level of impairment in a single model. The hypothesis is that the level of impairment (Incapacity, Disability, Discomfort) also scales on the logarithm of the concentration. The fit of the model to the data is surprisingly good, considering the large amount of random variation to be expected in estimates with the dagree of uncertainty expressed by the panel.

The model can be used to simplify additional Delphi studies of other pollutants. Thus, with the model, it is necessary to obtain estimates from panelists only of the base case, and the remaining cases can be predicted by the model. Similarly, if empirical (experimental) data can be obtained for one of the cases, the model can be employed to furnish a rough extrapolation technique to extend the data to other conditions.

The degree of impairment scale was defined independently for the different

pollutants and different populations. What the model indicates is that the respondents translated the specific symptoms into a more general scale which applied to all cases. This suggests that the scale can be extended to include a much wider range of degrees of impairment than the three points included in the present study. The formulation of a more extensive scale would allow extension of the dose-response model to include a wider range of phenomena. One immediate benefit would be the possibility of using data (e.g., lethality or life-threatening cases) not applicable to the present scale.

III. PSYCHONUMERIC HYPOTHESIS

One new outcome of the present study is an hypothesis concerning the way in which individuals make estimates of uncertain quantities. The data suggest strongly that individuals treat uncertain numbers as if they were on a logarithmic scale, rather than on the ordinary arithmetic scale. This assumption, which can be called the psychonumeric hypothesis in analogy with the better known psychophysical phenomena, rationalizes the observed lognormality of the distributions of estimates, and explains the experimentally observed exponential increase in error with the size of the number being estimated.

However, the hypothesis poses significant questions concerning the appropriate measure of error in formulating pollution criteria. Roughly speaking, the significance of a given error for potential health effects is perceived by the panel of experts as being related to the logarithm of the error, not to the error in terms of concentration levels. Again, roughly speaking, concentration at levels appropriate for CO, an error of 100 ppm is not perceived as 10 times as serious as an error of 10 ppm, but only as twice as serious (using logarithms to the base 10).

The question whether the psychonumeric phenomenon should be taken into account in formulating policy is an issue that exceeds the terms of reference of the present study.

SECTION 2

CONCLUSIONS

The following conclusions are based on the additional analysis of Delphi data presented in the 1975 EPA report on "Human Health Damages from Mobile Source Air Pollution".

1. Distributions of individual estimates for air pollution concentrations which may cause damages to human health are approximately log-normal.

2. The relationship between variance and geometric mean of individual estimates is compatible with the psychonumeric hypothesis.

3. Insufficient data exist in the HHD study to add to the assessment of self-ratings.

4. Estimated confidence ranges (upper and lower limits) are probably large underestimates of the actual confidence limits.

5. A relatively simple model of the concentration as a function of percent of population affected and degree of impairment fits the data rather well.

Some implications of these results for Delphi studies in the air pollution area are: (1) The conclusion of previous studies that the standard deviation of the distributions of estimates is a valid indicator of the accuracy of the geometric mean is compatible with the HHD data; however, some care must be taken in applying this rule to estimates which are systematic variants of the same estimate. (2) The value of obtaining estimates of expected range (upper and lower limits) appears marginal. (3) Several possibilities for extending the scope of estimation models by subjective scales (e.g., of degree of impair-

ment, or of relative severity of an illness) look promising. (4) The psychonumeric phenomenon poses serious issues concerning the role of judgment and the notion of error in decision problems.

SECTION 3

RECOMMENDATIONS

1. A simple dose-response model has been developed by this study. This model opens the possibility of formulating a scale of degree of impairment which would cover a much wider range of symptom states than those embodied in the three categories of conditions used in this study, Incapacity, Disability and Discomfort. A similar study should be undertaken by policy-making organizations such as EPA to further refine this model. Generalization of this model to fit in more comprehensive scales should not be a large step.

2. There are numerous decisions made by governmental agencies based on judgments. The delphi techniques used in this study have essentially provided a statistical basis of using judgments for decision-making. The resolution to the question of how should a policy organization such as EPA use or react to the delphi data may involve a detailed analysis of the organization's decision-making procedures.

3. There are some fundamental areas in the Delphi methodology requiring some in-depth explorations. These areas include calibration of standard deviations, self-ratings, estimated confidence and psychonumeric hypothesis. Investigation of these fundamental areas should be undertaken by agencies such as National Science Foundation.

SECTION 4

INTRODUCTION

This project was first initiated in June, 1973 and completed in August, 1975[1], as an in-house effort by the California Air Resources Board (CARB) with funding support from the U. S. Environmental Protection Agency (EPA). At the request of Dr. John Jaksch, the EPA contract officer, the project was extended in order to include more comprehensive data analysis.

The initial analysis of the data generated by this study of dose-response relationships for air pollutants was incomplete. The panel judgments obtained in the study constitute one of the best and most complete sets of data that have been collected in a Delphi study of health effects, and it fully warrants the additional analysis presented in this report. The data analysis reported in Section 7 is focused on two issues: (a) a more complete and rigorous treatment of the reliability of the data; (b) a more complete examination of the information contained in the data, e.g., concerning the interrelationships between type of pollutant, type of impairments and population type.

The conceptual framework within which the analysis of the HHD data is based upon is presented in Section 5. The concepts discussed in Section 5 include theory of errors and group judgment, expected error, the psychonumeric hypothesis, self-evaluation and estimated confidence ranges.

SECTION 5

CONCEPTUAL BACKGROUND

A. Theory of Errors and Group Judgment. The basic approach to estimation that is applied in the following analyses could be called the theory of errors. There are other theoretical approaches to numerical estimation that could be taken, e.g., treating the judgments as probabilistic statements, or treating them as expressions of an underlying model in the minds of the respondents[2]. However, the theory of errors is the simplest theoretical structure that can be applied to data of the type generated in the HHD study, and introduces the fewest assumptions.

Given a set of responses X_i to a numerical question, where $i = 1,\ldots,n$ indexes the individual members of a group, the theory of errors assumes that each individual response X_i is a function of three components, the true answer T, a bias term B_i, and a random error term R_i. Ordinarily, the function is assumed to be additive, i.e.,

$$X_i = T + B_i + R_i \qquad (1)$$

The bias term is presumed to be a function of the specific question, and of the information which the individual has concerning the question. The precise relationship between the bias term and amount of information is obscured by the difficulty in defining amount of information for the case of an expert answering an uncertain technical question. The assumption that the size of the bias term is inversely related to the amount of information appears plausible in a qualitative way.

The random error term is also considered to be a function of the specific question and of the individual's information. But in addition, it is a

random variable. If the question is repeated, T and B_i are constants, but R_i will vary in a random way. In practice, direct observation of this variation on repetition of a specific question is inhibited by memory effects, changes of information (if the repetitions are separated by significant periods of time), and by other more obscure factors such as degree of attention, motivation, and the like. In order to measure the random error, it is usually necessary to examine the responses to a large set of questions, and assume that a common mechanism is generating the deviations from the true value. For this reason, R_i is often called the "residual error" or "unexplained variation."

Despite the fact that individual distributions of responses to specific questions are usually not available, it is possible to make inferences concerning the nature of these distributions from data concerning sets of responses generated by a group of individuals, providing the responses are independent. This point will be discussed further in sub section 3.

It is usually assumed that the random error is sufficiently characterized by its mean, which by definition is zero, and by its dispersion or standard deviation. However, for many investigations, it is further assumed that the random error is normally distributed. This more restricted assumption will be made for most of the analyses which follow. The elementary theory of errors model, then, is equivalent to asserting that the individual "selects" his response out of a distribution that is normal around some mean that is displaced by the bias, B_i, from the true response T, as illustrated in Figure 1.

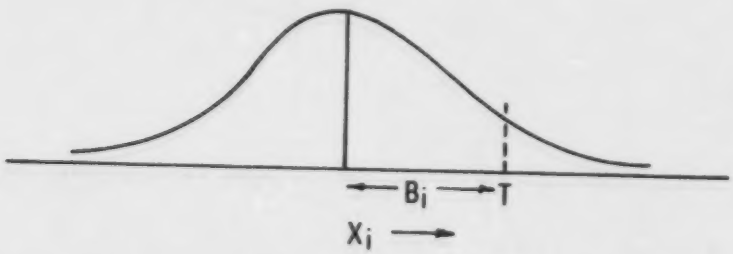


FIG. 1 ILLUSTRATIVE DISTRIBUTION OF INDIVIDUAL RESPONSES

The notion of bias has not received as much attention in psychological literature as the notion of random error (they are often lumped together as "error"); a simple illustration from a physical situation may make the idea clearer. Suppose there is a marksman firing at a target who has not compensated adequately for windage or distance. His pattern of shots might look like the dots in Figure 2, which are clustered about a point displaced from the center of the target. The displacement illustrated by the solid line in the figure is the bias of the pattern; the offset from the center of the pattern, illustrated by the dashed line, is the random error of the specific shot labelled X. It should be clear from this illustration that the notions of bias and random error are idealizations. The "biassing influences", such as wind and adjustment for distance, which are assumed to be constant throughout the trial.

FIG. 2 ILLUSTRATION OF BIAS AND RANDOM ERROR

Referring back to Figure 1, if the bias is unknown, the process appears to be a random selection of a response X_i out of a distribution around a mean M_i where $M_i = T + B_i$. In Figure 1, B_i is negative.

An additional consideration arises in dealing with free responses to numerical questions, i.e., responses which are not limited to a few categories. The distribution of random errors is usually not normal when plotted against the quantity being estimated, but is skewed in the direction of increase of the quantity. In part this arises from the fact that the quantities being estimated have a natural zero but no natural upper bound. However, a more basic psychological process appears to be operative here that will be discussed at greater length on page 21 in the section on psychonumeric scaling. For the moment, it appears to be the case empirically that over a wide variety of types of questions, the distribution of responses is not normal; rather, the distribution of the logarithms of the responses is normal. In technical terms, the responses are lognormally distributed.

Lower case letters will be used to designate the logarithms of the corresponding capitalized terms; thus $x_i = \log X_i$, $b_i = \log B_i$, $r_i = \log R_i$, etc. Lower case letters will designate the observed (sometimes called the sample) statistics of logarithmic quantities, m for the observed mean, s for the observed standard deviation. Thus $m = 1/n \sum_i x_i = 1/n \sum_i \log X_i$, and $s^2 = 1/n \sum_i (x_i - m)^2$, where the X_i are the observed responses of a group. Small Greek letters will be used to designate the theoretical parameters of distributions of logarithmic quantities; in particular μ designates the mean of a distribution of log quantities, and σ its standard deviation.

Rewriting (1) in logarithmic form, we have

$$x_i = t + b_i + r_i \qquad (2)$$

Note that (2) is equivalent to asserting

$$X_i = T\, B_i\, R_i \qquad (3)$$

Figure 3 shows the distribution of responses of upper-class and graduate student subjects to several hundred general information type questions, about 5000 responses in all*. To construct Figure 3, the responses to each question were normalized by setting $z_i + (x_i - m)/s$. The figure shows the density distribution of e^{Zi}. As is evident from inspection, the fit of the lognormal approximation is very good.

If we set $\mu_i = t + b_i$, the assumption of lognormality implies that the density distribution of individual i's responses, $D_i(x)$ has the form

$$D_i(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_i} e^{-\frac{(x - \mu_i)^2}{2\sigma_i^2}} \tag{4}$$

The corresponding distribution for the non-transformed estimate X_i is

$$D_i(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_i} \frac{1}{X} e^{-\frac{(\log x - \mu_i)^2}{2\sigma_i^2}} \tag{5}$$

Given (4) or (5) and the assumption that individual responses are independent, the distribution of the geometric mean G of a set of responses can be computed. The assumption of independence of responses, at least on the first round, is one of the basic features of the Delphi process, and is the justification for the rule of anonymity.

Setting $G = \{\pi X_i\}^{\frac{1}{n}}$ then $m = \log G$. The distribution, $D(m)$ is then

$$D(m) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{(m - \mu)^2}{2\sigma^2}} \tag{6}$$

where $\mu = 1/n \sum_i \mu_i$ and $\sigma^2 = 1/n^2 \sum_i \sigma_i^2$. Formula (6) is derived using theorem 2.3 in Aitcheson and Brown.[3]

---

* The set of experiments included studies at the Rand Corporation and at the Center for Computer-Based Studies at UCLA. They will be referred to collectively hereafter as the Rand studies and the data generated by the experiments will be called the Rand data[4,5].

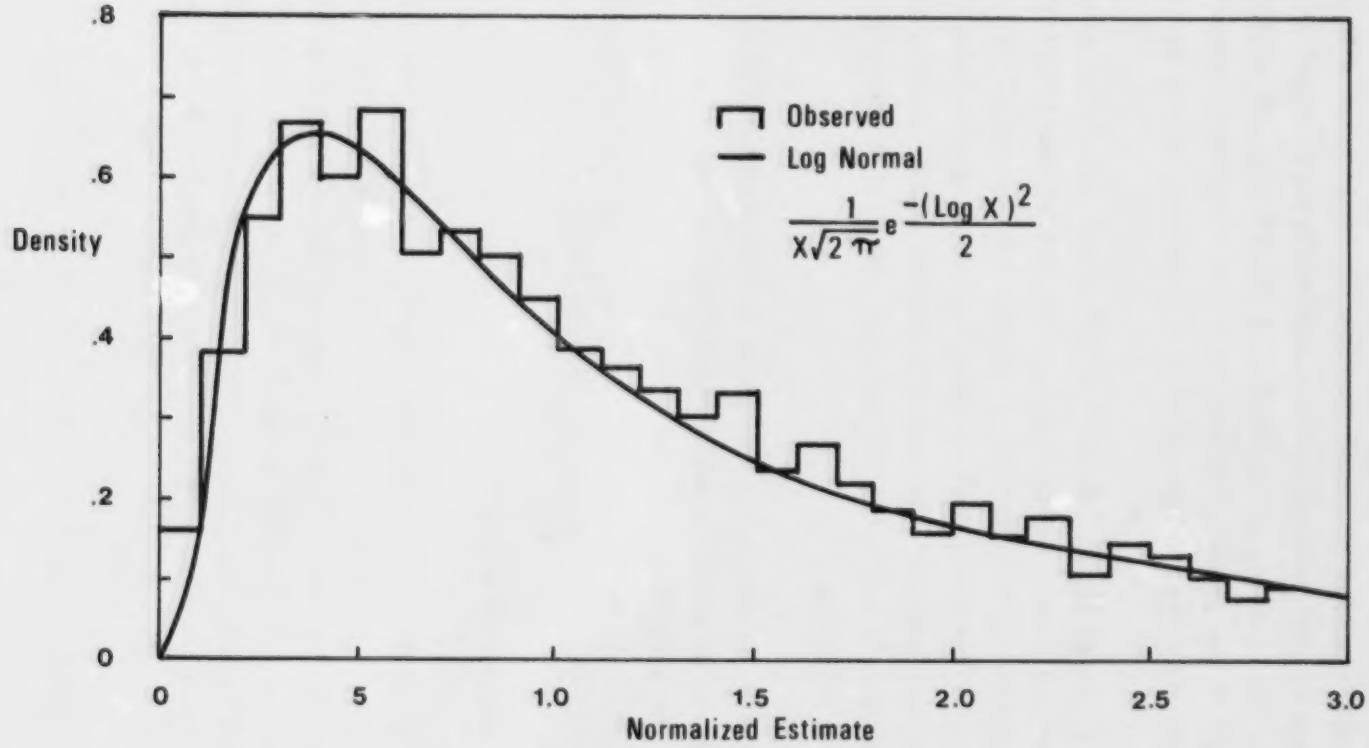


FIG. 3 DISTRIBUTION OF INITIAL ANSWERS

(6) states that the geometric mean of a set of independently lognormally distributed responses is itself lognormally distributed with a mean equal to the average of the individual means, and variance σ^2, equal to 1/n of the average of the individual variances. A specific set of responses, then, is a sample out of the joint distribution of the individual responses, and the sample mean m is the maximum likelihood estimator of the group mean.[4]

B. Expected Error. Formula (6) displays one of the advantages of group over individual estimation; namely, the standard deviation of the group mean is smaller by a factor of $1/\sqrt{n}$ than the average of the standard deviations of the individuals. This feature of group estimation--well known in statistical sampling theory--is a precise way of stating that the process of combining many estimates "washes out" the random variability of the individual estimates.

If the distribution of the group mean were unbiassed, $\mu = t$, then the only error to be expected would arise from the residual variability of the mean; that is, the expectation of the error $|m - t|$ can be computed as

$$\text{Ex } |m - t| = \sqrt{\frac{2}{\pi}}\,\sigma = \sqrt{\frac{2}{\pi}}\,\frac{\sqrt{1/n \sum_i \sigma_i^2}}{\sqrt{n}} \qquad (7)$$

where Ex designates expectation.

The assumption $\mu = t$ is equivalent to the assumption that the average bias $1/n \sum_i b_i = 0$. This assumption is generally not fulfilled in experimental studies of estimation. Figure 4 displays the relationship between observed average error and observed standard deviation for the Rand experiments. The upper line is the observed relationship (least squares fit) while the lower line is the relationship that would be expected for $\mu = t$. Random error accounts for only about 1/3 of the total error. Thus

$$\text{Total expected squared error} = b^2 + \text{Ex } (r^2) \qquad (8)$$

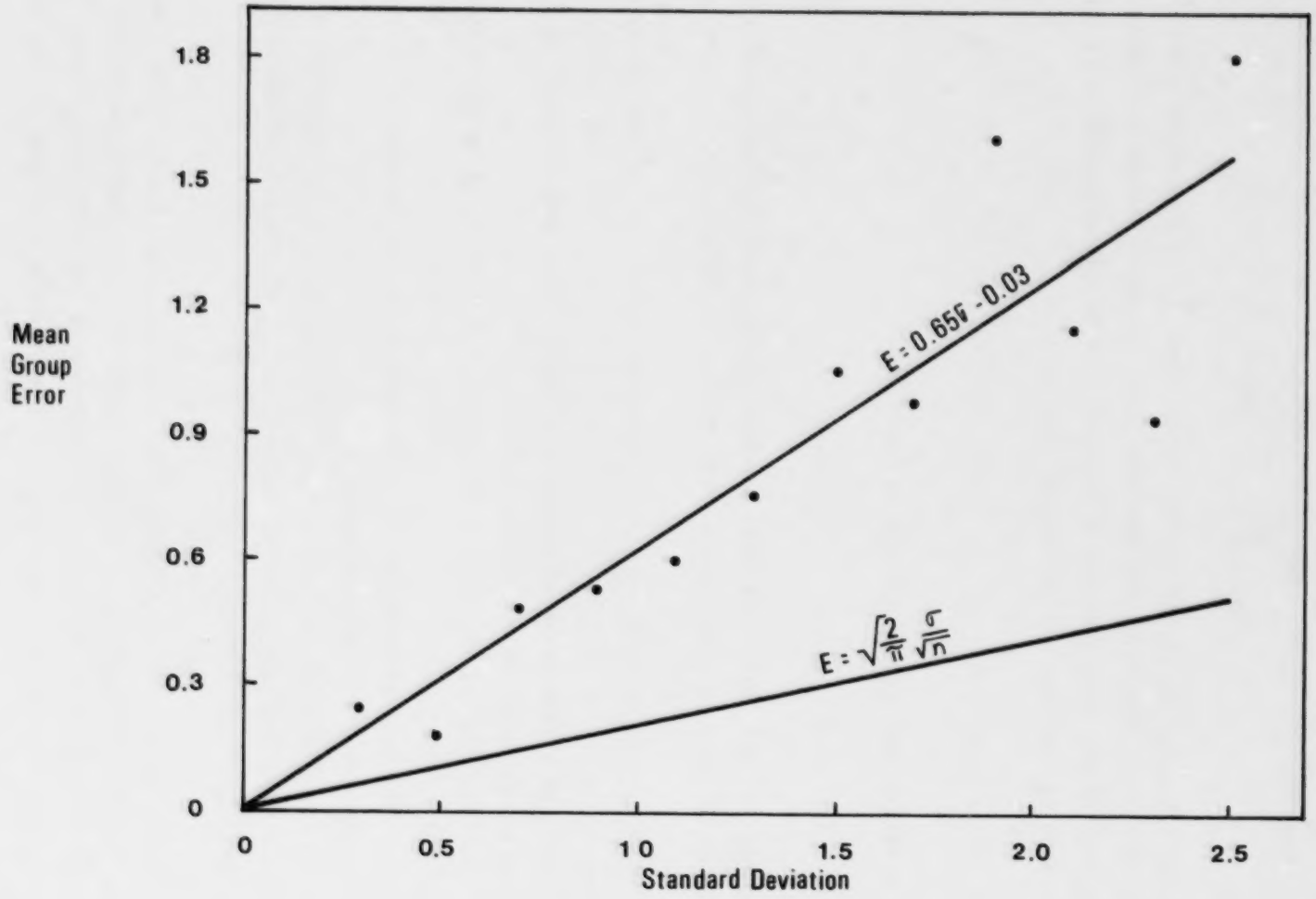


FIG. 4 INVARIANCE OF E / σ

where $b = 1/n \sum_i b_i$ = bias of the mean, and $r = 1/n \sum_i r_i$ = random error of the mean. Equation (7) states that

$$Ex(r) = \sqrt{\frac{2}{\pi}} \quad \frac{\sqrt{1/n_i \Sigma \sigma_i}}{\sqrt{n}}$$

Thus, as n becomes large, the random error term in (8) approaches zero. There is no such guarantee for the bias term. However, there is a weaker guarantee for the bias term which displays a second advantage of group over individual estimation. It is easy to show that

$$b^2 \leq 1/n \sum_i b_i^2 \tag{9}$$

With a little more trouble it can be shown that

$$|b| \leq 1/n \sum_i |b_i| \tag{10}$$

Equations (9) and (10) state that the bias of the mean is always less than, or at worst equal to, the average bias of the individuals.[7]

Thus we can conclude from this elementary exposition of the theory of errors:

(a) The random error of the mean is always less than or equal to the average random error of the individuals.

(b) The bias of the mean is always less than or equal to the average bias of the individuals.

(c) The average error of the mean computed as in formula (7) is a lower bound for the expected average error.

(d) Confidence limits (e.g., the 95% confidence range expressed as $\pm$ 2s) computed from the observed standard deviation are a lower bound for expected confidence limits.

If the empirical relationship between average error and standard deviation shown in Figure 4 has general validity, then the lower bounds expressed in (c) and (d) are weak; the larger contribution to the total error comes from bias. Figure 4 indicates that bias and the standard deviation of the random

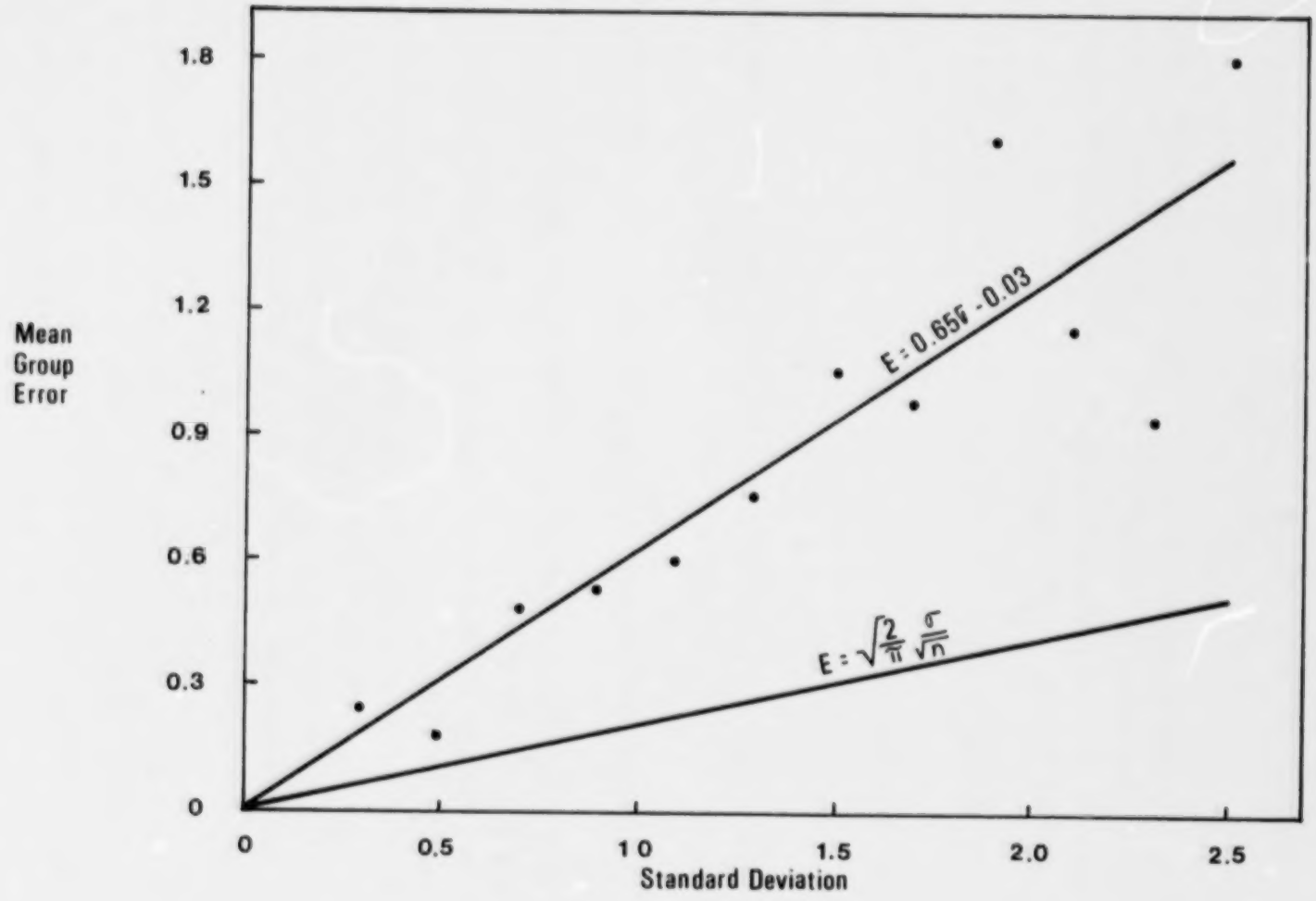


FIG. 4 INVARIANCE OF E / σ

error are coupled, that is, as the standard deviation increases, the bias increases. This is a plausible coupling on the assumption that both random variation and bias have a single cause, namely incomplete information. However, there is no theory at the present time that specifies the nature of this coupling. Hence, the only way to estimate the bias in the HHD data is to assume that the empirical relationship holds for estimates generated by a panel of doctors on medical topics as well as for the estimates generated by upper level college students in the Rand studies[4,5] with topics of general information.

C. The Psychonumeric Hypothesis. The formulas in the two preceeding sections were derived on the assumption that the individual distributions are lognormal. As remarked above, these individual distributions are not usually observable directly. However, it is possible to give a fairly substantial grounding to the assumption. Aitcheson and Brown have shown that if the distribution of m (the observed mean) is lognormal, then the underlying individual distributions are also lognormal. If there were no bias, i.e., if $\mu = t$, then the assumption that the observed set of means all come from a common family of distributions could be tested by forming the statistic $z_m = (m - t)/s$ and testing the distribution of z_m for normality.

Figure 5 shows the cumulative distribution of z_m from the Rand data in a log probability diagram. The straight line is the cumulative distribution predicted for a lognormal with $\mu = 1/3$ and $\sigma = 1/3$. The fit is relatively good. Figure 6 shows the same curve plotted as a density distribution. The curve from Figure 3 is plotted on the same scale for comparison. The existence of bias is evident. Rather than $\mu = 0$ (no bias), $\mu = 1/3$. However, the lognormality is well demonstrated, and the inference that the underlying individual distributions are lognormal but biassed appears reasonable.

The lognormal distribution for individual estimates can be derived from the central limit theorem and the assumption that the residual error can be decomposed into many small errors that combine multiplicatively. However, without some justification for the multiplicative combination, that route appears highly ad hoc.

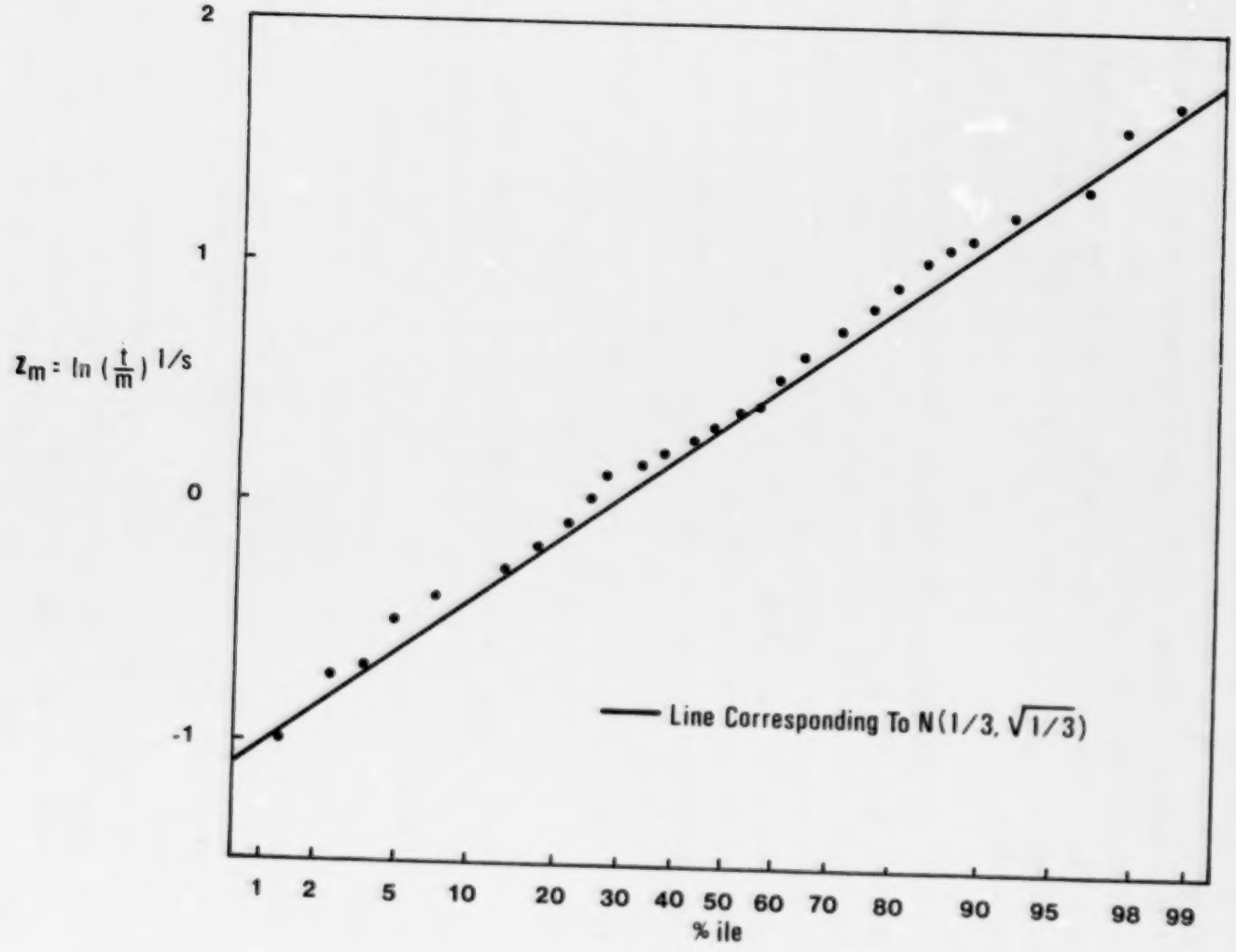


FIG. 5 CUMULATIVE FREQUENCIES OF z_m ON PROBABILITY SCALE

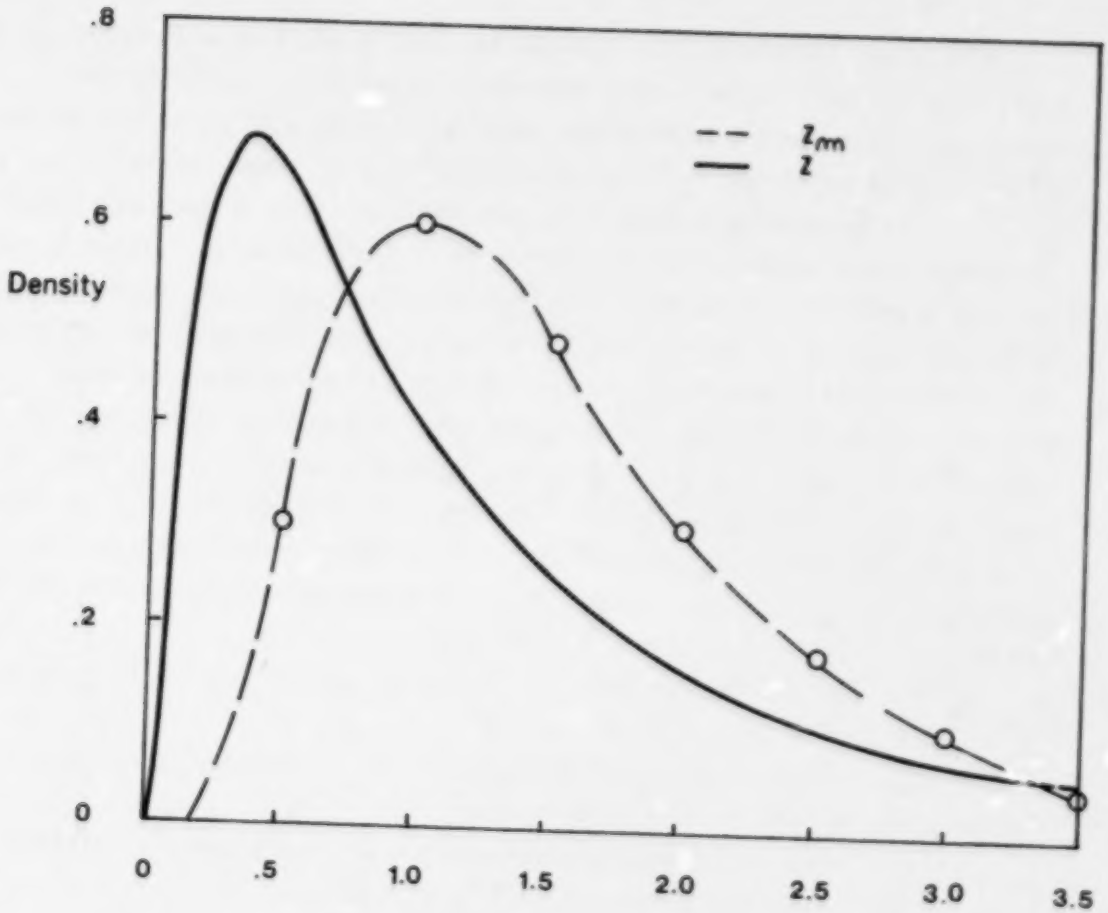


FIG. 6 DENSITY DISTRIBUTION OF e^{z_m}

A more general approach that seems to be more in line with a large body of psychological work is to postulate a non-linear transformation between physical magnitudes and the psychological scales on which these magnitudes are estimated. Formula (2) above would be interpreted as asserting that for the individual making an estimate, it is not the usual real number metric which is "even" (additive), but rather the logarithmic transform of the real number metric.

Non-linear transformations between physical quantities and psychological magnitudes are well-known in many perceptual modalities, including the logarithmic relationship between the physical intensity of sound and psychological intensity that has given rise to the decibel scale of sound intensity. There has been a long-standing debate as to the "general" form of such psychophysical transformations, whether they are basically logarithmic as postulated by Weber-Fechner, a power law as claimed by S. S. Stevens, or some more general form[8]. To my knowledge, the question whether a similar type of transformation exists for numbers generated by more general types of estimation such as those generated in non-perceptual estimates of uncertain quantities has not been treated in the psychological literature. However, there is now a fairly large amount of data which suggests that the numbers which "come to mind" in attempting to answer uncertain questions also have a non-linear relationship to the corresponding physical quantities. These data are summarized in figures listed below:

1. The lognormality of observed distributions in Figure 3 and Figure 12* on page 43.
2. The lognormality of the distribution of the means of group estimates in Figure 5 and Figure 6.
3. The linear relationship between log error and standard deviation of log estimates in Figure 4.
4. The non-linear scaling of standard deviations with size of the estimate in Figure 9.
5. The non-linear scaling of error with size of estimate in Figure 10.
6. The logarithmic distribution of first digits of responses to numerical questions.

---

* In Fig. 3 the distribution is plotted against e^{z_m} whereas in Fig. 12 it is plotted against z_m. Both graphs indicate a normal distribution on the logarithm of the responses.

The sixth item is somewhat out of the context of the others, and needs some background. One of the intriguing features of statistical tables such as are found in almanacs and similar reference works, is the distribution of the first digits of the numbers. There is a tendency for the first digits to be distributed in a logarithmic pattern; specifically the frequency of digit d (d = 1,...,9) is roughly proportional to log(d + 1) - log d.[9] A somewhat more general hypothesis would be that the numbers x are themselves distributed as 1/x. This would imply that not only the first digits but the second and subsequent digits would also have the appropriate distribution. For example, the frequency of d as a second digit would be proportional to

$$\sum_{i=1}^{10} \left[\log(10i + d + 1) - \log (10i + d)\right]$$

This hypothesis has not been verified in detail. A relatively large body of data would be needed to generate stable statistics, but a quick try of a few thousand numbers selected more or less at random out of an almanac looks favorable. The distribution of second digits is shown in Figure 7. The dotted lines shows the theoretically expected frequencies.

More relevant to estimation, when several thousand responses to estimation questions were analyzed in terms of the distribution of first digits, they exhibited precisely the same logarithmic distribution as the data from the almanac tables. The distribution of the first digits of the responses is given in Figure 8. The only major departure from the theoretical distribution is an evident preference for the digit 5.

One might be tempted to believe that the distribution of first digits in the responses is being "driven" by the corresponding distribution in the true answers which the subjects are trying to approximate. Indeed, the true answers exhibited the logarithmic distribution. However, the two distributions were completely independent. Whatever psychological mechanism generated the distribution of responses, it was independent of the mechanism that generates a logarithmic distribution of first digits in the almanac tables.

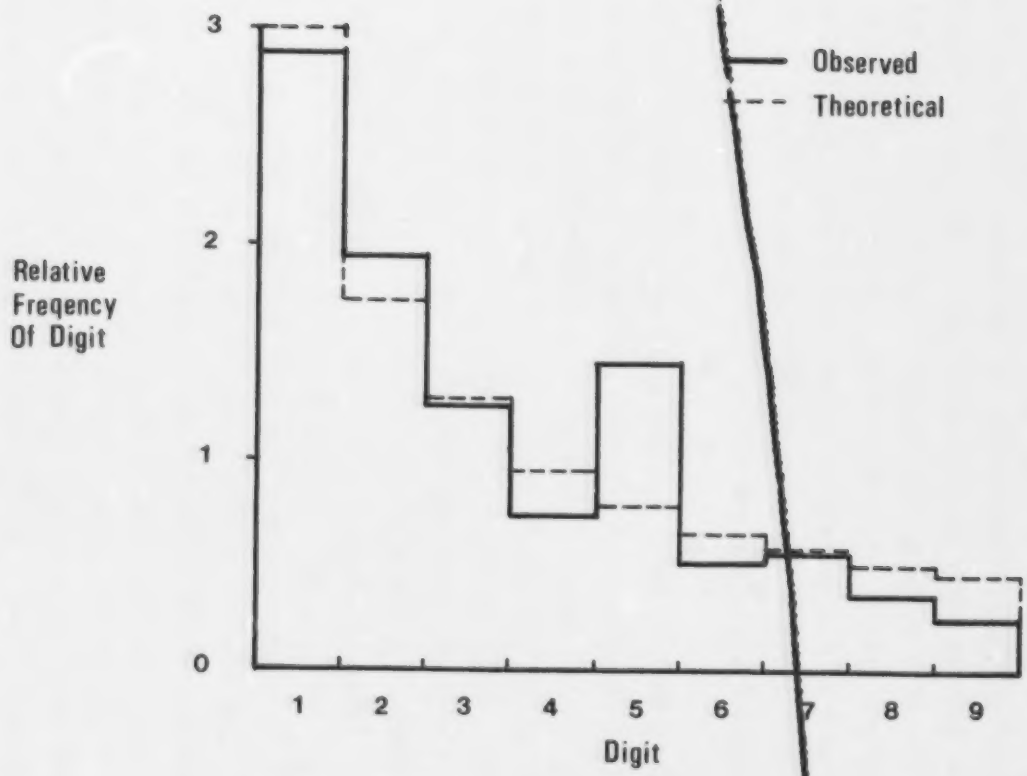


FIG. 8 DISTRIBUTION OF FIRST DIGITS, SUBJECT RESPONSES (5,037 RESPONSES)

The distribution of first digits, then, is partial confirmation of the hypothesis that the real number system in the minds of the respondents, at least for the task of estimating uncertain numbers, is distributed like 1/x. Further confirmation of the hypothesis comes from an examination of the scaling of the standard deviation with the size of the number being estimated. Figure 9 is a plot of the average observed standard deviation against the logarithm of the true answer to the question. Figure 10 is a corresponding plot of average error against log true. Both figures show an increase in the respective indices with increasing size of the estimates. It is interesting to note that the peculiar dip in the vicinity of 10^6 in both figures appears to be an artifact introduced by the particular choice of questions for these experiments, but no specific property of the questions has been identified that would explain the anomaly.

Both the log error and the standard deviation of the log estimates are invariant under a multiplicative scaling of the estimated quantity. Thus, if $\{X_i\}$ is the set of estimates for one question, with a true answer T, and $\{aX_i\}$ and aT are the corresponding responses and true answer for another question, then the standard deviations of the log responses are the same for the two sets of estimates, and the error $|m - t|$ is the same. This is easily seen for the error since $1/n \sum_i |\log aX_i - \log aT| = 1/n \sum_i |X_i + \log a - \log T - \log a| = 1/n \sum_i |\log X_i - \log T|$. Figures 9 and 10 show immediately that the subjects were not scaling their estimates in this simple multiplicative manner. Roughly speaking, the data indicate that scaling is being carried out at the level of the logs, not at the level of the original estimates. Put in other terms, if estimates are generated for two questions, one of which has an answer T and the other an answer $T' = aT$, then the responses $x_i' = \log X_i'$ will be scaled like $x_i \log a$, not as $x_i + \log a$. A theoretical scaling of error based on this hypothesis is plotted as the dashed line in Figure 10. The fit is obscured somewhat by the erratic nature of the data in the vicinity of 10^6, but otherwise the hypothesis that the subjects are scaling their responses on the log appears to fit the data quite well.

The psychonumeric hypothesis applied to dose-response estimates, for example, is equivalent to stating that the relative significance of an additional

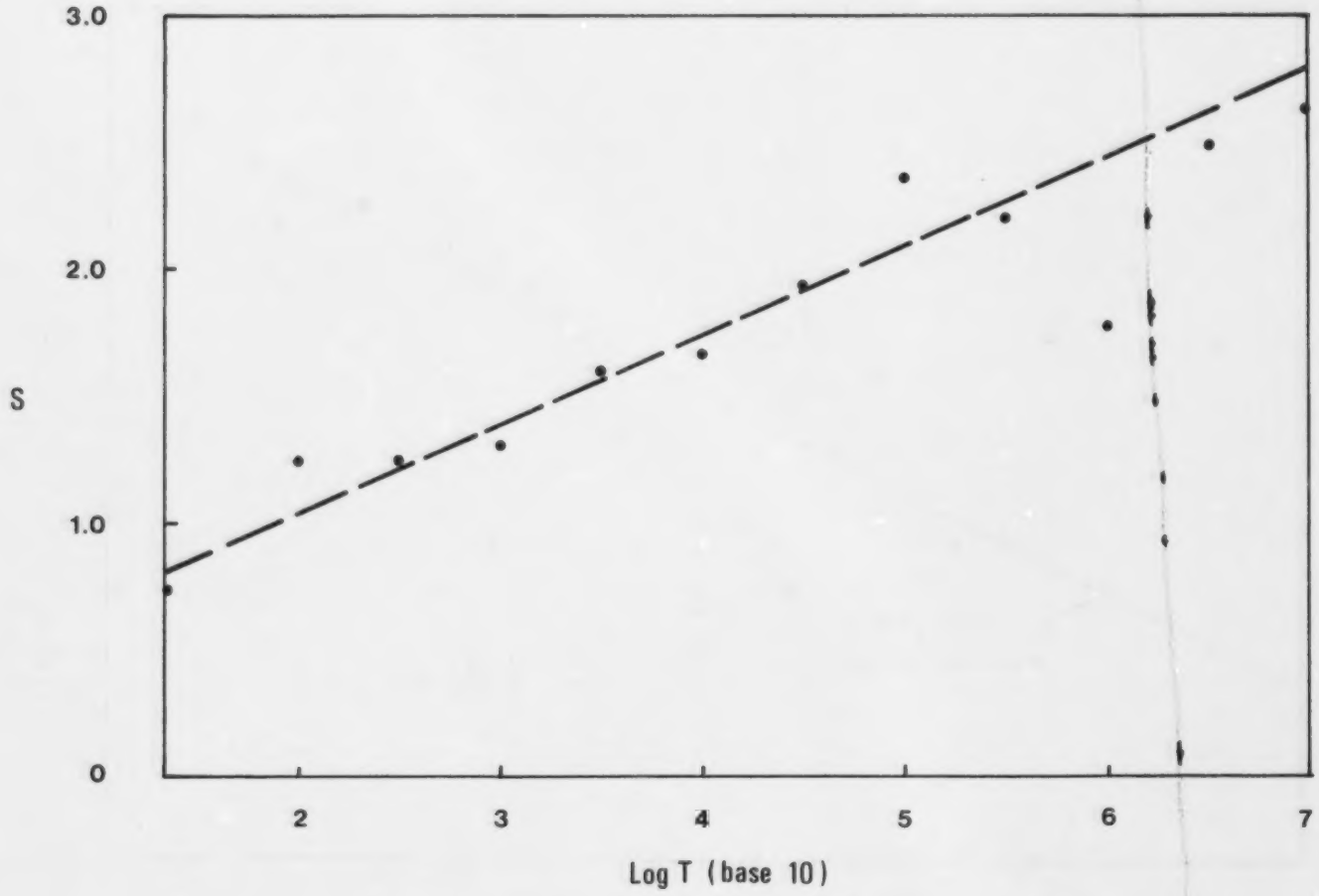


FIG. 9 AVERAGE STANDARD DEVIATION AS A FUNCTION OF LOG TRUE

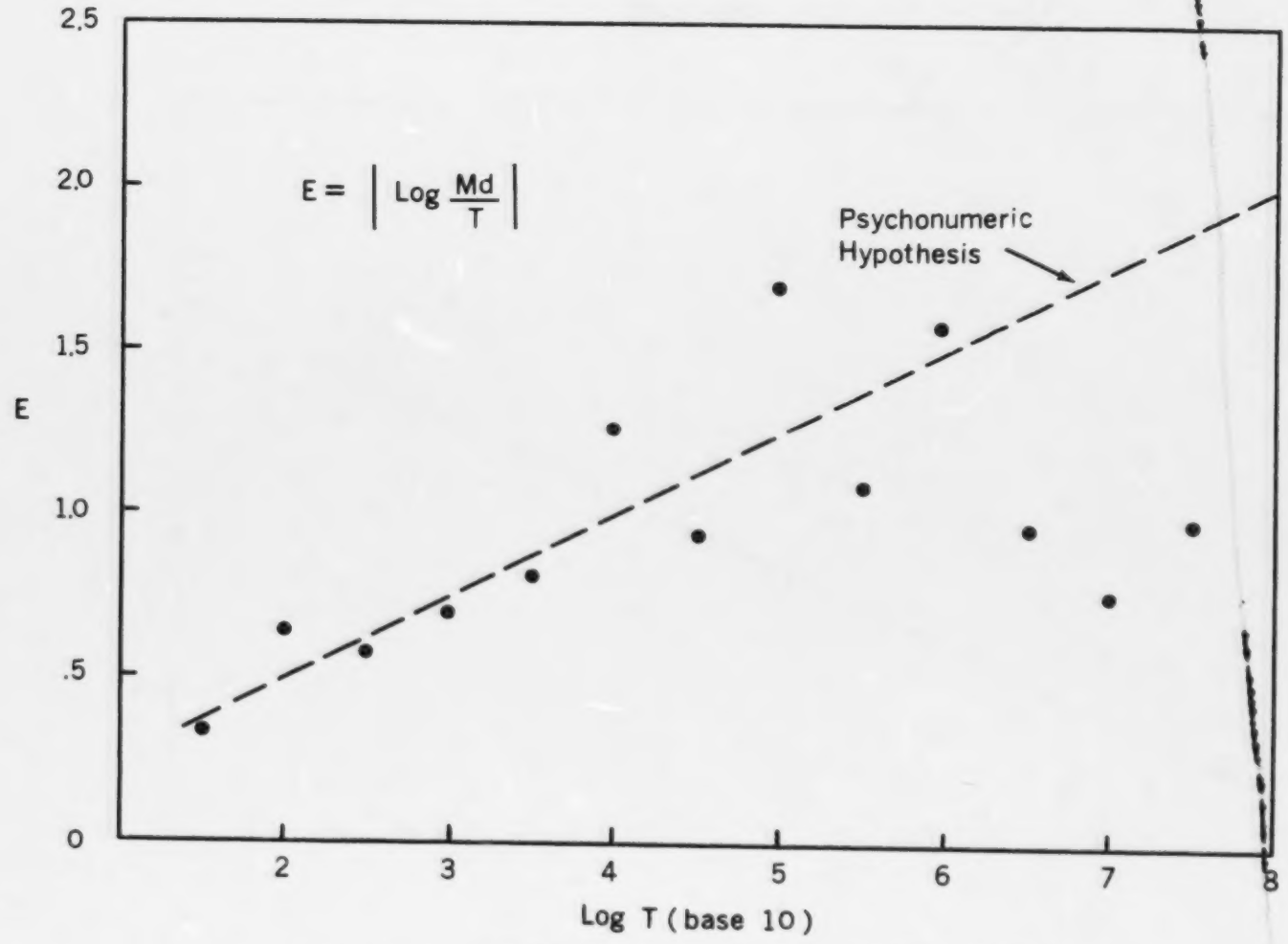


FIG 10. AVERAGE ERROR AS A FUNCTION OF LOG TRUE

increase in dosage depends on the percentage increase, not on the absolute increase. At the level of significant effects for Oxidant, about 0.8 ppm for 50% disability for a normal population according to the panel, an increase in concentration of 0.2 ppm raises the expected percent of population affected to 90%. 0.2 ppm is a 25% increase over 0.8 ppm. For CO, the estimated dosage for Normal, 50% population, disability, is 170 ppm. An increase of 0.2 ppm is an increase of 0.12%, which the panel clearly finds inconsequential.

Thus from several different directions, the lognormality of group and individual distributions, the distribution of first digits in estimates, and the scaling of distributions, with size of estimates, the same conclusion is arrived at, namely, the psychonumeric hypothesis that subjects think of the real numbers in terms of the log transformation.

One reason for emphasizing the psychonumeric relationship is the consequences of such a non-linear transformation for the notion of error. For example, if an individual is asked to estimate the intensity of a sound in terms of psychological units (technically, in sones) and makes an error of, say, 10% at the 100 decibel level (he should have said 100 sones, but said 90) he has made an error of a factor of 10 in the physical intensity of the sound. On the other hand, if he makes an error of 10% at the 30 decibel level (he says 27 instead of 30) the error in the physical magnitude is "only" a factor of 2. More generally, if we express the percent error in the psychological magnitude Ψ as

$$\frac{\Psi - \Psi(T)}{\Psi(T)} = e = \frac{\log M - \log T}{\log T}$$

Then $M = T^{e+1}$. If an acceptable error for a highly uncertain number is, say 30%, then at the level of $T \sim 10^6$, an "acceptable" response would be $(10^6)^{1.3} = 10^{7.8}$, an error of 6000 percent.

Scientists who are accustomed to thinking of error in terms of physical magnitudes would probably be very unhappy to switch to defining error in terms of a psychological magnitude. Actually, the situation is not serious if the relevant phenomena are related to the psychological magnitude. In

the case of the intensity of sound scale, for purposes of estimating noise pollution, where this is defined in terms of psychological responses, the decibel scale is quite appropriate. If there were reason for believing that physiological responses to air pollution were related to the logarithm of the concentration, rather than to the concentration directly, then there would be a happy match between the psychological estimation process and the relevant physiological processes.

As it turns out, the first approximation model of the HHD data relating percent population affected and degree of impairment to dosage is logarithmic. Whether this is a sufficient reason to relax criteria for error concerning health damage estimates is a decision which the medical community may want to take seriously.

D. Self-Evaluation. An additional indicator of accuracy can be obtained by asking each respondent to rate his estimates. Such self-ratings have been used routinely in Delphi exercises, and respondents do not appear to have any particular difficulty in making the judgments. In the past these ratings have not been tied to any particular theory of estimation and their use for assessment of the solidity of group judgments has been largely informal.

In practice, each respondent has been left a good deal of freedom in interpreting what the self-rating entails. In the Rand experiments, each respondent rated each of his answers (on the first round) on a scale of 1 to 5, where 1 meant "sheer guess" and 5 meant "I know the answer". The correlation between individual error and individual self-ratings was about -.25, large enough to indicate an association, but not large enough to use the self-rating as a figure of merit for individual answers. On the other hand, the correlation between average self-ratings and error of the median was about -.60, a respectable degree of association[5].

The correlation between average self-rating and observed standard deviation was also quite high, about -.67. The correlation between standard deviation and error was about .63. A combined index of average self-rating and standard deviation gave a multiple correlation of .67 with error. The improvement of the combination over the standard deviation alone is small, a result of the high correlation between the two components. Figure 11 shows

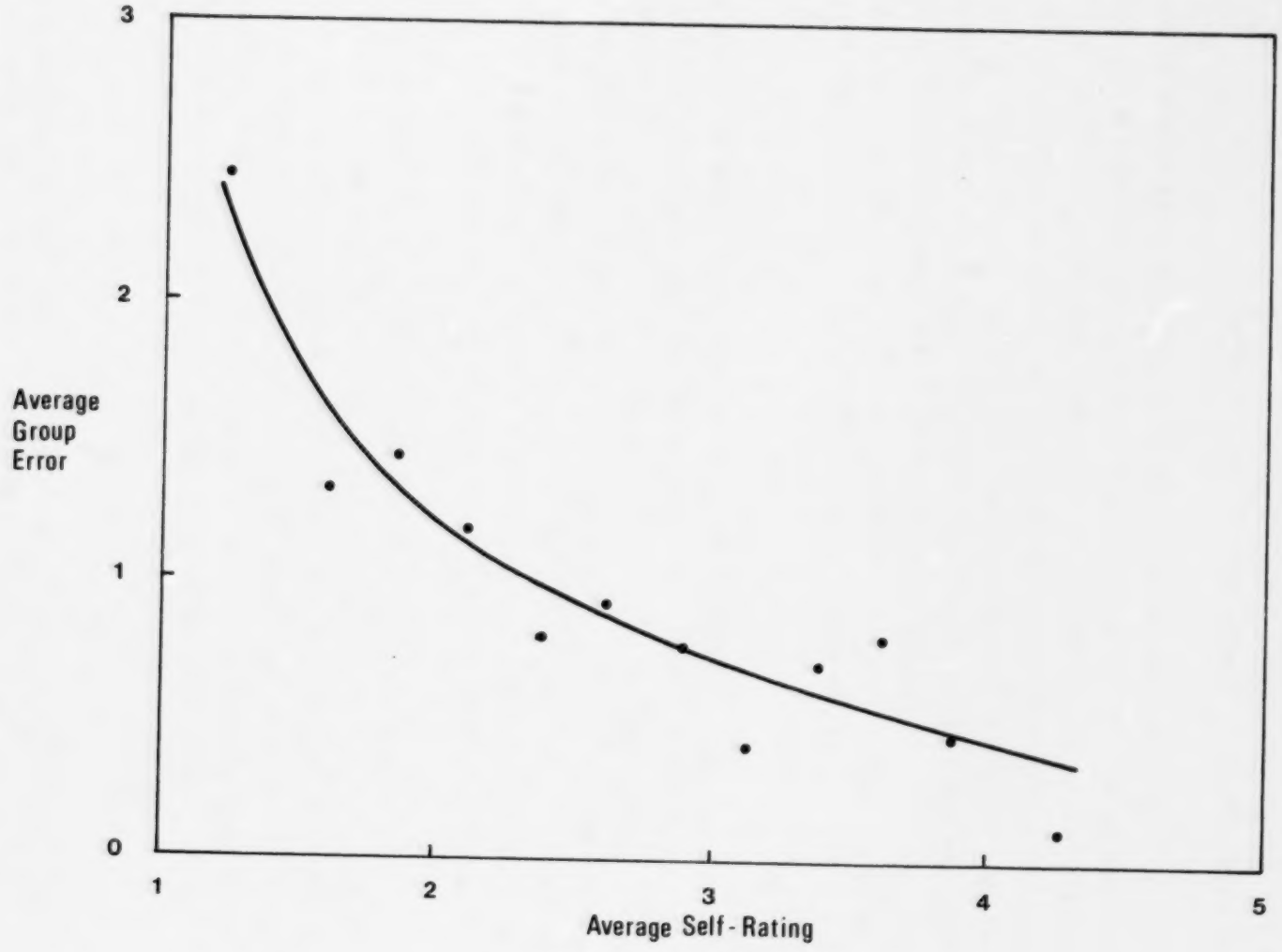


FIG. 11 GROUP SELF-RATING

a plot of average group error (average error of the median) against average self ratings for a large subset of the Rand data[6].

It is difficult to integrate the self-rating with the theory of errors outlined above. If the self-rating expresses mainly the uncertainty that is exemplified in the variability of individual responses, then it is of limited value in assessing group responses, since the standard deviation is a more reliable measure of the variability. The high correlation between standard deviation and self-ratings in the Rand experiments would tend to suggest that the self-ratings are expressing roughly the same thing as the variability. However, a number of studies have suggested that both bias and variability are a function of the amount of knowledge that the individual has concerning the given question[10].

E. Estimated Condifence Ranges. One additional index of certainty was included in the HHD study. Each respondent was asked to give a high and a low estimate, where these were interpreted roughly as the 95% and 5% confidence limits on the respondent's "best estimate". To include these estimates in the theoretical approach would require treating the responses as probability judgments, with the best estimate interpreted, e.g., as the individual's geometric mean for his probability distribution.

The foundations for this more extensive theory of estimation have been developed[7]. However, it dod not appear worthwhile including it in the present analysis of the HHD data. In other studies where such limits have been elicited, and where objective answers to the questions were available, the bounds have proved to be almost uniformly too narrow[11]. Capen concludes from his studies that, however the bounds are characterized, what is obtained is something more like a 40% to 50% range, rather than a 90% range*.

---

* The terms "limit" and "range" can be confusing. The 95% confidence limit refers to the value, X_{95}, of the quantity where 95% of the cases are expected to be lower than this value. The 5% confidence limit, then is the value of X for which 5% of the cases are expected to be lower. The 90% confidence range refers to the interval within which 90% of the cases are expected to fall. The 90% range is generally taken to be the interval between the 5% and 95% confidence limits.

SECTION 6

RESEARCH METHODS

The previous section contains the conceptual framework within which the analysis of the HHD data will be pursued. The very general form of theory of errors can, of course, be applied to practically any replicated data. However, to make the results of previous studies applicable to the HHD data, it is necessary to show a sufficiently close analogy so that it is plausible to assume that the same type of estimation processes were involved in the previous studies and in the HHD study. In effect this comes down to determining whether the distributions are roughly lognormal, and whether the estimates exhibit the psychonumeric form of scaling.

The HHD study interrogated a panel of 14 medical experts concerning the dosage (concentration with one hour exposure) of three different air pollutants required to produce three levels of severity of symptoms in 14 population groups. These population groups are:

Normal
Children
Old Age
Heart Condition, Mild
Heart Condition, Severe
Hay Fever, Sinusitis
Influenza
Upper Respiratory Infection
Asthma
Acute Viral Bronchitis
Acute Bacterial Pneumonia
Chronic Respiratory Diseases
Mild Chronic Obstructive Lung Disease
Severe Chronic Obstructive Lung Disease

The dosage estimates were obtained for 0%, 10%, 50% and 90% of the respective populations. This led to 504 separate estimates by each panelist. In addition, each respondent was asked to estimate the upper and lower limits for each concentration estimate, to estimate the duration (in hours) of the resulting symptoms, and to express a self rating on a scale of 1 to 10. Thus each panelist generated 2520 judgments, 35,280 estimates in all.

The exercise was iterated one additional round. Estimates which dod not fulfill a criterion of "good agreement" (interquartile range $\leq$ 1/2 median) were resubmitted to the panel, along with a report of the median and interquartile range from the first round. The estimates from this second round, along with the carry-over of the estimates that were not iterated, constituted the final outputs of the study.

The HHD study is one of the most thorough Delphi studies involving professional respondents and dealing with a significant substantive topic that has been conducted to date. During the study there was insufficient time to examine the data for the light it could shed on a number of basic issues relevant to the assessment of the validity of the panel estimates. Validity assessment were based on the results of previous theoretical and experimental studies involving different types of respondents and different kinds of subject matter.

The following analyses are an attempt to investigate the analogy between the HHD data and the experimental more completely. The analysis is based on the first round estimates. Primarily this is because the theory of errors assumes that the estimates are independent, and the feedback of first round results during iteration leads to non-independent estimates on the second round. The feedback step leads to convergence (reduction in standard deviation) and also leads to a small but significant increase in the accuracy of the estimates[6]. However, the convergence is relatively much greater than the error reduction. The ratio of error to standard deviation roughly doubled between round one and round two for the experimental

study cited. Thus the first round standard deviation is a more diagnostic measure of the variability than the second round standard deviation. It should be noted that this is true only if no additional information beyond the results of the first round is introduced at the second round. This condition was fulfilled in the HHD study.

The validity assessments for the HHD study were based on (1) the experimentally observed relationship between standard deviation of the log estimates and the log error as shown in Figure 4 and (2) an experimentally observed relationship between self-ratings and error, Figure 11. These experimental results were obtained using university upper-class and graduate students as subjects and questions of general information such as might be contained in an almanac or statistical abstracts as subject matter. A recent study of professional petroleum engineers and analogous general information questions shows that there is no significant difference in types of results obtained as a function of the type of respondent.[12] In addition, analysis of one type of professional estimate, namely bids on oil leases, shows a close similarity between such judgments and the student estimates for almanac questions, namely, lognormality of responses and variances appropriate to the assumption of log scaling.[13] However, these results were not examined in terms of group judgment.

SECTION 7

RESULTS

The first subsection below takes up some of the conclusions that can be drawn directly from the theory of errors. This means basically what can be concluded omitting the possibility of bias. The second subsection examines the lognormality of the responses in the HHD study. The third looks at the scaling of standard deviation as a function of the size of the estimates. Subsection four deals with the correlation between the three indices of uncertainty, standard deviation, self-rating, and estimated confidence range. The final subsection takes up the possibility of formulating a model of the estimates relating dosage to degree of impairment and percent population.

A. Theory of Errors. As noted above, the average error AE computed from the standard deviation of the mean is a lower bound to the expected error. Similarly, confidence limits computed from the standard deviation of the mean furnish a lower bound for the expected confidence range. The criterion for acceptability (on the first round) imposed in the original HHD report, namely S $\leq$1/2 median is roughly equivalent to s $\leq$.5. This implies a standard deviation of the mean $s_m \leq .5/\sqrt{14} = .134$. Estimates which pass this criterion would have an AE $\leq$.106 (Formula (7)) which implies an average error in the geometric mean of $\pm$ 11% (Additive error in the log estimates translates into multiplicative error in the original estimates.) If the usual confidence limits is defined as $\pm 2s_m$, the confidence limits on the geometric mean would be $\pm$ 31%.

Whether these criteria for expected error and confidence limits are "acceptable" (taking into account the fact that they are lower bounds) would probably depend more on the type of decision being contemplated than on statistical logic.

Turning to the HHD data, the majority of estimates for oxidant (OX) fit the criterion of $s \leq .5$. None of the nitrogen dioxide (NO_2) or carbon monoxide (CO) estimates fit the criterion. The fact that a small percentage of the NO_2 and CO estimates fit the criterion in the form $S \leq 1/2$ median can be ascribed to discrepancies between the median and the geometric mean, and to discrepancies between S as observed and S as it would be computed from s. The second type of discrepancy perhaps requires some explanation. Assuming that the distributions are lognormal, the appropriate maximum likelihood estimators for the distribution parameters are m for the logarithm of the geometric mean and s for the standard deviation of the log estimates. The geometric mean is then estimated by e^m and the standard deviation by the formula

$$S^{*2} = e^{2m+s^2}(e^{s^2}-1) \tag{11}$$

Let S designate the standard deviation computed directly from the estimates, $S^2 = 1/n \sum_i (X_i - M)^2$, $M = 1/n \sum_i X_i$. M, of course, is not the geometric mean. In the HHD study, the median was used as a surrogate for the geometric mean and the observed standard deviation as a surrogate for S^*, the former because earlier studies had indicated that the median is slightly more accurate than the geometric mean, and the latter because, on the basis of examining only the OX data, S appeared to be a sufficiently good approximation to S^*. That was a too hasty conclusion, as indicated by Table 1 which displays the ratios S^*/S and GM/Md for the Normal population. I, Da, Dc, are abbreviations for Incapacity, Disability, Discomfort, respectively. The very large discrepancy $S^*/S = 2.94$ for OX, 0%, Dc should be disregarded. One of the estimates for this case is 0, and S^* is highly sensitive to the approximation for log 0 (which theoretically = $-\infty$). Otherwise, for OX, the median is a relatively good approximation to the geometric mean, and S^* is relatively close to S. However, for NO_2 and CO, this is not the case.

Table 1 suggests that for the HHD data, assessments should be based on s and m, not on the median and S. That is, the best estimate should be defined as the geometric mean e^m, and the estimated standard deviation should be S^*, the standard deviation computed by (11).

This conclusion has some consequences for the evaluations presented in the

TABLE 1. The Ratios of S*/S and GM/Md for the Normal Population

Pollutant	% Population	S*/S Impairment			GM/Md Impairment		
		I	Da	Dc	I	Da	Dc
OX	0	.98	1.10	2.94	.97	1.10	.90
	10	1.01	1.07	1.01	1.05	1.14	.97
	50	.89	.89	1.04	1.07	1.10	.90
	90	.73	.81	1.13	1.07	1.12	.89
NO_2	0	.97	1.32	1.65	1.09	.82	.89
	10	.76	.88	.91	1.31	1.12	1.03
	50	.95	.92	1.19	1.32	1.04	.78
	90	1.10	.77	1.35	1.69	1.26	.88
CO	0	.84	.83	.77	1.30	1.11	.95
	10	1.11	.90	.94	1.45	1.17	.97
	50	1.10	.83	.88	1.45	1.20	.82
	90	1.05	.87	1.02	1.61	1.23	.94

I = Incapacity

Da = Disability

Dc = Discomfort

original HHD study. For example, as noted above, none of the NO_2 and CO estimates fulfill the criterion $s \leq .5$. The criterion $s \leq .5$ has a certain amount of arbitrariness associated with it. It was selected in part because of the fact that estimates which fulfill the criterion do not (on the average) improve with iteration[1]. The question whether this criterion needs revision in the light of HHD data will be taken up in the conclusion section.

As an example of the application of the "pure" (assumption of no bias) theory of errors to the HHD data, Table 2 displays the average s (averaged over all cases) for the three pollutants, and the average error AE and associated confidence limits CL computed from s, AE computed from (7), CL = $\pm$ 2s.

TABLE 2. Average Error and Confidence Limits from Theory of Errors

Pollutant	Average s	AE	CL	AE'	CL'
OX	.489	10%	$\pm$28%	37%	$\pm$ 60%
NO_2	.908	21%	$\pm$63%	80%	$\pm$238%
CO	.795	18%	$\pm$53%	68%	$\pm$215%

Table 2 also shows, for comparison, the AE and CL values (primed entries) based on the empirical relationship between error and s (Figure 4). In this case, CL' = $\pm$ 2s + b, and b = AE' - AE.

Again, whether these average AE and CL figures are acceptable (keeping in mind they are lower bounds) would presumably depend upon the decision problem.

B. Lognormality. Figure 12 shows the frequency distribution of z scores for all "best estimate" responses of the panel. The z score is defined as

$$z_i = \frac{x_i - m}{s}$$

where x_i is an individual log estimate, m is the mean of the log estimates, and s is the observed standard deviation of the log estimates. Computation of

z scores were made for each set of 14 responses to a given pollutant, degree of impairment, percent population case; hence there were 504 separate distributions treated. The average of these 504 distributions is shown in Figure 12, where the ordinate indicates the relative frequency with which the z score falls within the intervals indicated on the abscissa. If the distributions tend to be log-normal, then the distribution in Figure 12 should be approximately normal. The solid line indicates the observed distribution; the dotted lines show the expected normal distribution.

Altogether, there are 7056 estimates represented in Figure 12. If we test the hypothesis that each of these estimates was independently selected from a normal distribution, the observed distribution fails a test of goodness of fit at the .001 level, $\chi^2 = 230.8$ with 10 degrees of freedom. However, there is little justification for assuming that each estimate was independently selected. A distinction has to be drawn here between independence for estimates by different individuals, and independence for estimates by the same individual. Some care was taken to assure that estimates by different individuals were independent on the first round, and there is no evidence that the procedure was not effective. An entirely different issue is involved in the question whether individuals made independent estimates for each of the many cases they dealt with, or whether in effect estimates for closely related cases were generated by systematic modification of a single estimate.

We can reformulate the χ^2 statistic in the form

$$\chi^2 = n\left(\sum_i^m q_i^2/p_i - 1 \right) \qquad (12)$$

where n is the number of independent estimates, q_i is the observed relative frequency in cell i, p_i is the expected frequency in cell i, and m is the number of cells, m-1 is the number of degrees of freedom. The expression in the brackets can be interpreted as a measure of the degree of similarity between the observed distribution and the expected distribution, where 0 indicates perfect similarity. Define $\tau^2 = \sum_i q_i^2/p_i - 1$. In the case of Figure 12 $\tau^2 = .0327$, which indicates a high degree of similarity. However, the statistical significance of this measure depends on n. For $\tau^2 = .0327$, an n of 560 would be required

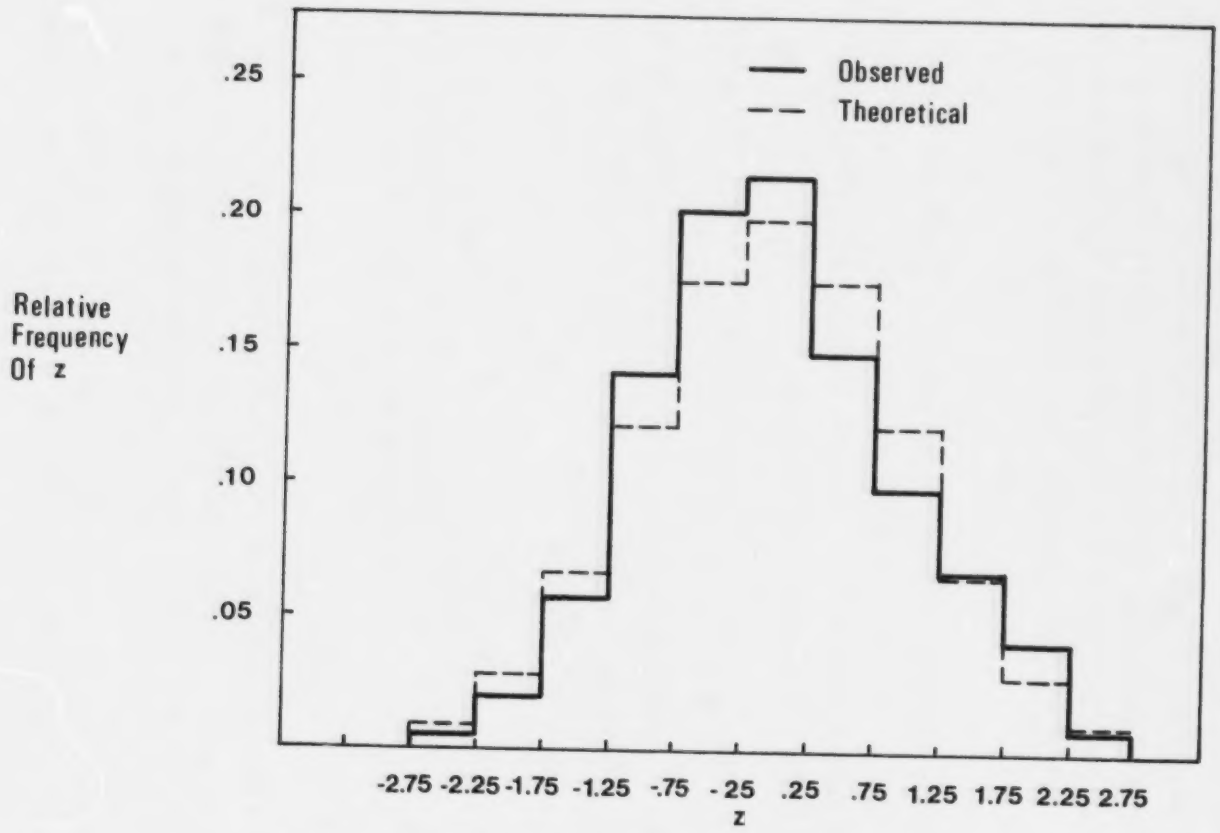


FIG. 12 FREQUENCY DISTRIBUTION OF z SCORES FOR ALL BEST ESTIMATES

to reject the hypothesis of normality at the .05 level.

It is difficult to specify a reasonable figure for n, the number of independent estimates. It is clear from the investigation of the model discussed below that there are highly systematic interrelations between the estimates for different percent population and degree of impairment cases. In addition there are strong correlations between individual estimates across pollutants and across populations, i.e., individuals who tend to give relatively low estimates in one case also tend to give relatively low estimates in others, etc.

A frequently employed measure of the degree of agreement within a set of judgments is Kendall's coefficient of concordance, W*. It is defined for rankings of a set of objects by a group of individuals. If R_{ij} is individual i's ranking of the j'th object, define

$$Q = \sum_j \left(\sum_i R_{ij} - \bar{\Sigma}\right)^2$$

where $\bar{\Sigma}$ is the average sum of ranks. Q is thus the sum of squared deviations of the sum of ranks from the average. The coefficient of concordance is defined as

$$W = Q/Q_{max}$$

where Q_{max} is Q computed for the case where all the respondents are in complete agreement on their rankings. W measures the divergence of the ranking from perfect agreement.

An approximate χ^2 can be derived for W by multiplying W by n(m-1) where n is the number of respondents and m the number of objects. In using the χ^2 tables, the number of degrees of freedom is taken to be m-1.

For example, consider the degree of agreement in the panel across populations with an otherwise fixed case. For this computation, the role of

---

* Kendall, M. G., Rank Correlation Methods, Charles Griffon & Co., London, 4th Ed., 1970.

"individual" and "object" is reversed. Each population generates a ranking of the individuals (the ranking of the individual estimates for the given population and case). There are thus 14 rankings of 14 "objects". For the case CO, 50%, Disability, W is .884, which gives a χ^2 of 160.9. With 13 degrees of freedom, W is significant well beyond the .001 level.

Another measure of agreement is the correlation coefficient. The product-moment correlation between estimates for Normal and Children populations for the case OX, 50%, Disability is .784, significant at well beyond the .01 level (two-tailed test).

The high degree of agreement between individual estimates across cases indicates that the estimates are not being selected at random out of a common distribution -- there is a high degree of dependence among the responses. With this in mind, it does not appear unreasonable to conclude that the data does not reject the hypothesis that the distributions are approximately log-normal.

On the other hand, it is also true that other distributions, such as a beta distribution, would fit the data equally well. Here we have to rely on the presumption borrowed from previous studies that the lognormal distribution is a likely candidate.

C. Logarithmic Scaling. As was stated in Section A.4, log-normal distributions are to be expected if the estimates are scaled on a logarithmic transform of the physical quantity being estimated. To that extent, Figure 12 gives some weight to the psychonumeric hypothesis. More direct evidence is furnished by Figure 13 which displays the observed standard deviation s as a function of the mean m for the Normal population. The dashed line is a continuation of the empirically derived relation between standard deviation and true answer displayed in Figure 9.

The data tends to lie along the empirical curve; however Figure 13 cannot be construed to establish the hypothesis of logarithmic scaling for the standard

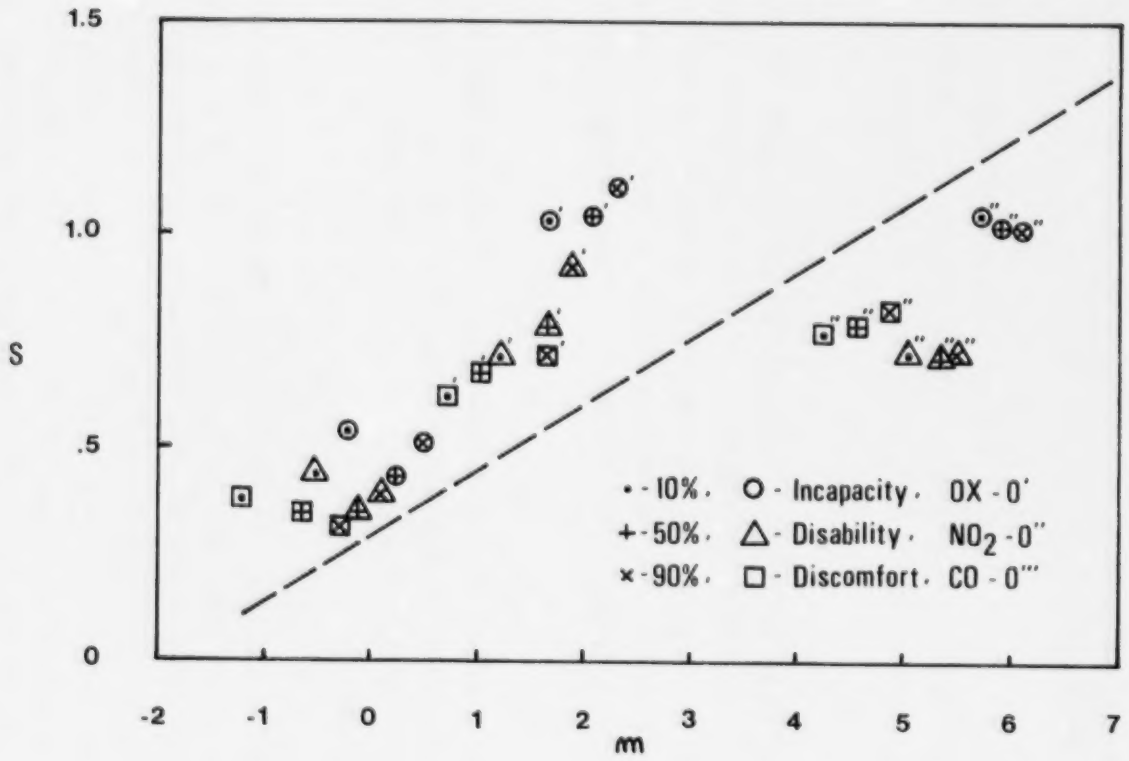


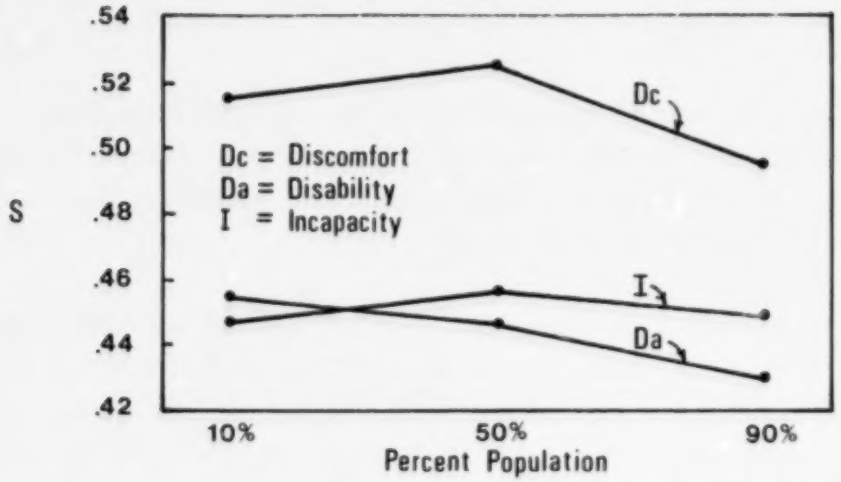
FIG. 13 OBSERVED STANDARD DEVIATION s AS A FUNCTION OF THE LOG GEOMETRIC MEAN m FOR THE NORMAL POPULATION

deviation. The data is somewhat sparse for this purpose. Each of the different pollutant cases form separate blocks. In addition, there are rather mysterious within block uniformities. In particular, the Incapacity cases for NO_2 and the Disability cases for CO seem "out of line".

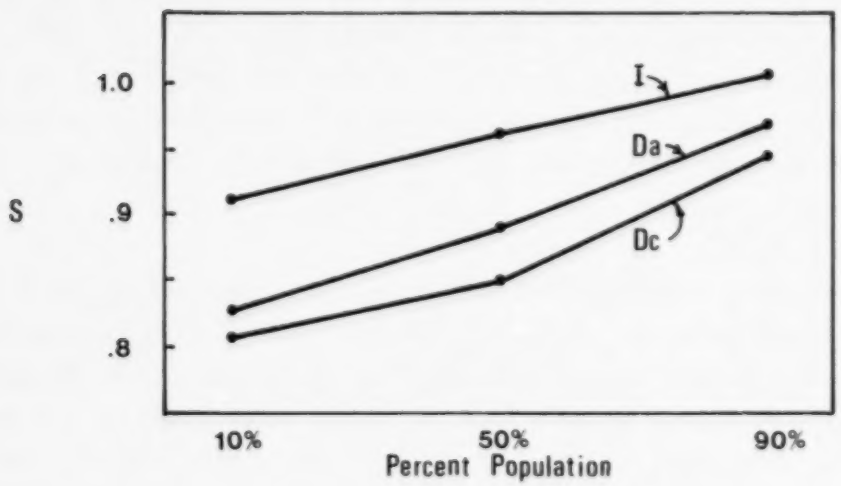
The average standard deviations for the various percent population and degree of impairment combinations are displayed in Table 3. In addition, the corresponding averages of the means are listed. Inspection of Table 3 shows a rather anomalous pattern of variation of s with percent population and degree of impairment. Only NO_2 exhibits the pattern that would be expected from scaling, namely increase of s with percent population, and increase of s from Dc to I. Figures 14 (a) - (c) show s plotted against percent population, and Figures 15 - 17 show s plotted against m, for the various pollutants. If the three pollutants showed similar anomalies it might be tempting to ask what generated them. As it stands, the anomalies are puzzling.

Summing up the picture on logarithmic scaling: Figure 13 is compatible with the hypothesis of logarithmic scaling, and taken in conjunction with lognormal distributions, is some evidence for assuming that the same general estimation processes are operative in professional judgments concerning dosage as in student responses to almanac questions. However, the data for variations within pollutant categories presents a rather muddled picture. The theory of errors in log form, and concomitantly the hypothesis of log scaling, is based on the assumption of independent estimates. The estimates concerning special cases within pollutants are clearly not independent. The implications of this lack of independence will be taken up after the investigation of models below.

A. OXIDANT



B. NITROGEN DIOXIDE



C. CARBON MONOXIDE

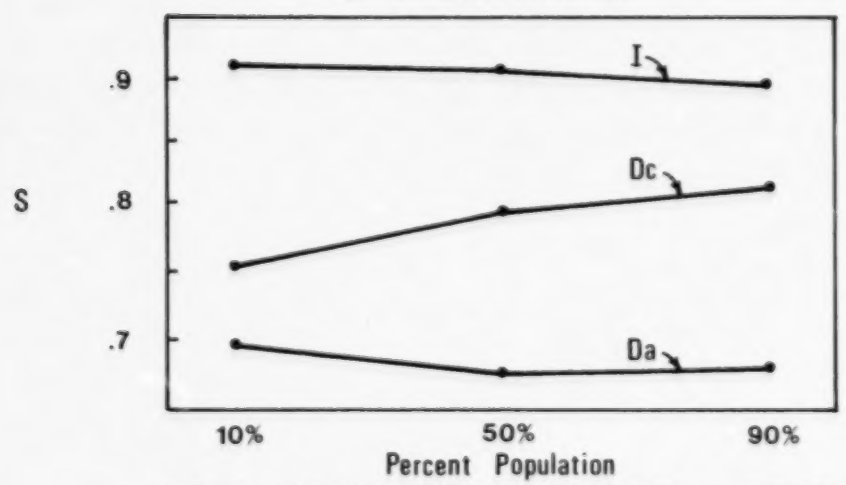


FIG. 14 AVERAGE STANDARD DEVIATION vs PERCENT POPULATION

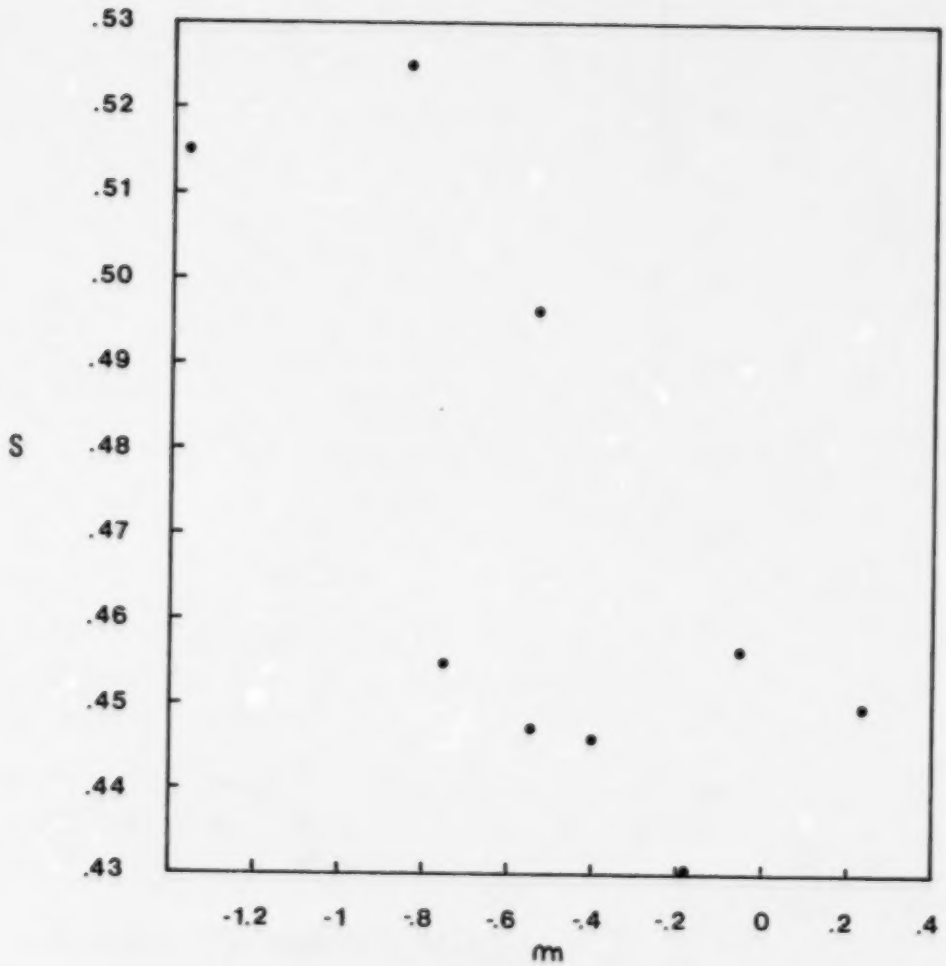


FIG. 15 AVERAGE STANDAND DEVIATION s vs AVERAGE MEAN m FOR OXIDANT

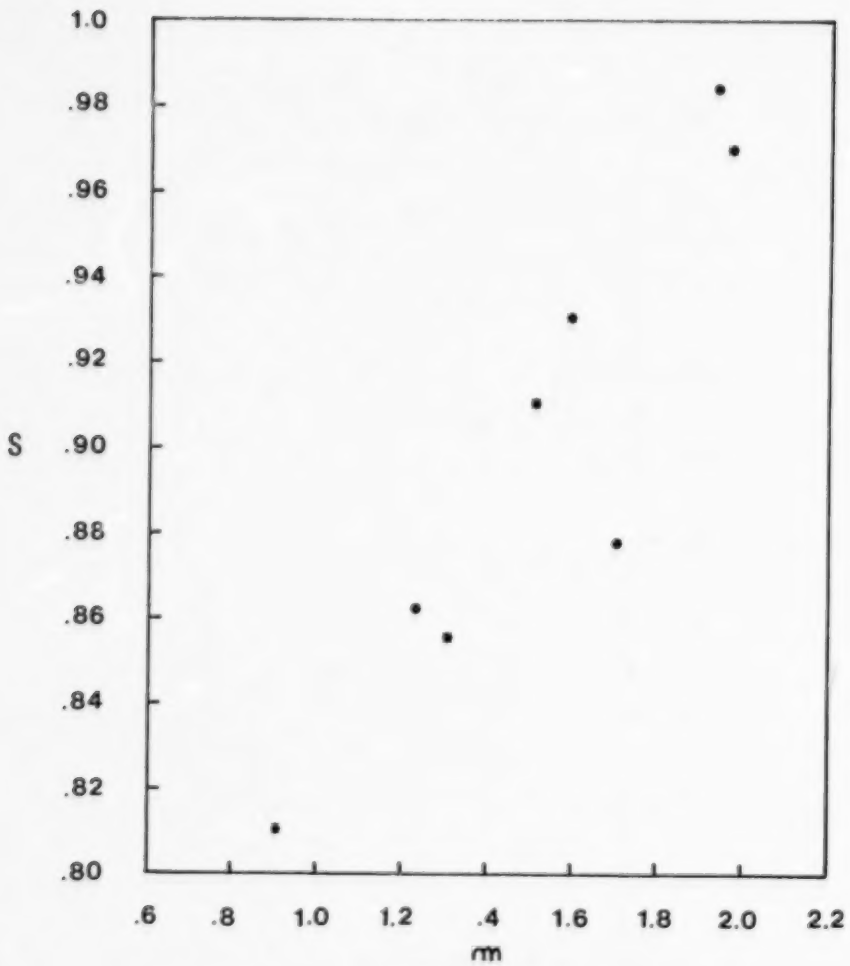


FIG. 16 AVERAGE STANDARD DEVIATION s vs
AVERAGE MEAN $\bar{m}$ FOR NITROGEN DIOXIDE

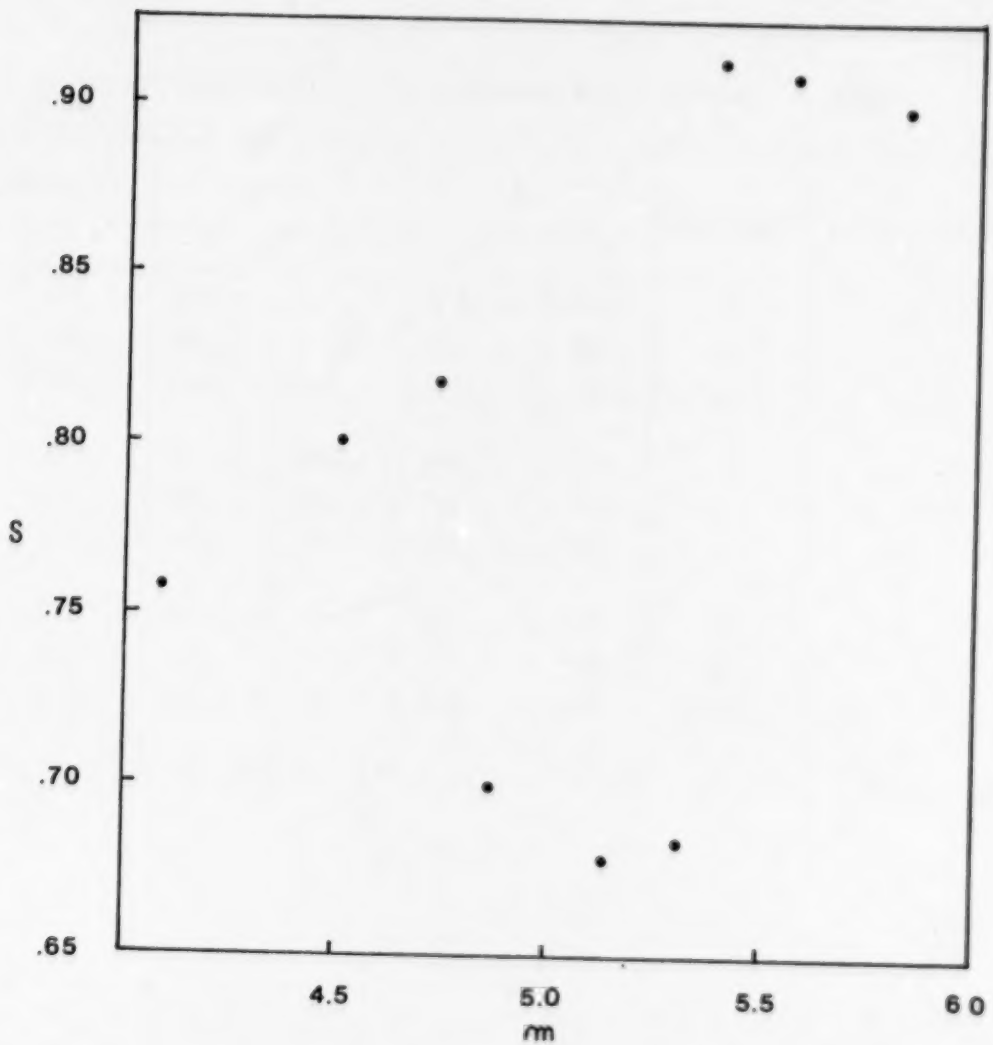


FIG. 17 AVERAGE STANDARD DEVIATION s vs AVERAGE MEAN m FOR CARBON MONOXIDE

TABLE 3: Average s and Average m of All Population Groups

% Population	Impairment	OX		NO_2		CO	
		s	m	s	m	s	m
10	I	.447	- .539	.910	1.480	.912	5.399
	Da	.454	- .751	.836	1.056	.699	4.874
	Dc	.515	-1.269	.809	.641	.757	4.192
50	I	.456	- .048	.958	1.755	.808	5.672
	Da	.446	- .395	.888	1.398	.678	5.151
	Dc	.525	- .824	.850	1.016	.800	4.511
90	I	.449	.236	1.011	2.089	.899	5.825
	Da	.430	- .189	.965	1.687	.682	5.314
	Dc	.496	- .531	.942	1.287	.819	4.741

D. Correlation of Indices of Unvertainty. The three measures, standard deviation, estimated confidence range, and self-rating, are all related to the amount of uncertainty the respondent has concerning a given estimate, and hence are related indirectly to the accuracy of the estimate. The standard deviation is related to accuracy by the theory of errors; the estimated confidence range and the self-rating are related by psychological assumptions concerning the perception on the part of the individual of the relative "solidity" of the evidence he has for his estimate. This psychological theory is not sufficiently advanced to make quantitative predictions concerning the relation between self-ratings or estimated confidence ranges and error. The empirical relationship for self-rating and error observed in the Rand studies was described above.

Since error could not be measured directly for the HHD data, the only analysis available was investigating the correlation among the three indices. The only pair out of the three for which correlations for individuals could be computed is confidence range and self-rating. Estimates were pooled across percent population to give 56 data points for each correlation; otherwise correlations were computed separately for each pollutant, degree of impairment, and population type. Thus 126 correlations were computed. These are displayed in Table 4.

Inspection of Table 4 shows that the correlations are uniformly rather small; the largest is -.244. Since R increases with certainty and Δy decreases, the correlations should be negative. About 1/3 of the entries in Table 4 have the opposite sign from what is expected. None of them reach statistical significance assuming 56 independent cases. At first sight, this appears to be dubious support for the hypothesis that both R and Δy measure the degree of certainty that the respondent has in his estimate. However, in pooling the responses across percent population, a significant reduction in the size of the correlations was introduced. This results from the fact that the average R is almost constant across the four subsets of data, but the average Δy declines sharply from 0% to 90%.

TABLE 4: Correlation Between Self-Rating and Confidence Range

Pollutant	Impairment	POPULATION TYPE													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
OX	I	.096	.111	- .077	- .001	- .079	- .012	- .151	- .164	- .018	- .085	- .028	- .117	- .123	- .080
	Da	.099	.090	- .028	.022	- .051	- .194	- .111	- .140	- .037	- .041	- .030	.032	.025	- .092
	Dc	- .0628	.128	- .091	.012	- .026	- .191	- .092	.012	- .084	- .002	- .042	- .009	- .032	- .078
NO_2	I	- .038	- .051	- .052	- .034	- .071	- .103	- .073	- .006	- .134	- .105	- .102	- .169	- .093	- .073
	Da	- .108	- .244	.062	- .063	- .024	- .066	.125	.001	- .094	.040	.168	.177	- .055	- .078
	Dc	.108	- .037	- .059	- .110	.047	- .067	- .066	.040	.028	.134	.243	.066	.026	.180
CO	I	- .093	- .117	- .066	- .050	- .063	- .071	- .051	- .053	- .006	- .058	- .058	- .053	- .031	- .094
	Da	- .024	- .051	.022	- .021	- .075	- .004	.013	- .027	.058	- .012	- .010	.012	- .010	- .033
	Dc	- .040	- .083	- .043	- .036	- .199	- .056	- .004	- .034	- .004	- .040	- .059	- .023	- .055	.049

I = Incapacity
Da = Disability
Dc - Discomfort

The correlation for pooled populations, r_{xy} is related to the correlations within the subpopulations $r_{x_i y_i}$ (where i indexes the subpopulations) by the formula

$$r_{xy} = \sum_i \left(\frac{n_i}{n}\right) \left(\frac{\sigma_{x_i}\ \sigma_{y_i}}{\sigma_x\ \sigma_y}\right) r_{x_i y_i} + \frac{\tilde{CV}\ (\bar{x}, \bar{y})}{\sigma_x\ \sigma_y} \qquad (13)$$

s_x and s_y are the standard deviations of x and y in the total population, s_{x_i} and s_{y_i} are the standard deviations in the subpopulations. $\tilde{CV}(x,y)$ is a generalized covariance of the means of the subpopulations where averaging is accomplished by the weights n_i/n. The total number of cases is designated as n, n_i the number of cases in subpopulation i. Because R is virtually constant, $\tilde{CV}(\bar{R}, \Delta\bar{y})$ is essentially 0. On the other hand, $s_{\Delta y}$ is uniformly larger than the $s_{\Delta y_i}$. Hence $r_{R\Delta y}$ will generally be much smaller than the $r_{R_i \Delta y_i}$. This effect is illustrated graphically in Figure 18. Here three subpopulations are shown, in each of which $r_{R_i \Delta y_i} = 1$. Obviously $r_{R\Delta y} \neq 1$. In fact it is .577 for the case illustrated in Figure 18.

This effect was checked by computing correlations separately for the 0%, 10%, 50%, and 90% subpopulations for the case NO_2, Children, Disability. The correlation with these four subpopulations pooled is .24. Table 5 lists the four subpopulation correlations separately. The correlations for 50% and 90% are significant at the .05 level (two-tailed test) for n = 14. The question whether the stringent two-tailed test should be used here is unclear, since the sign of the correlation is also a part of the hypothesis being tested.

There was insufficient time to rerun the correlations separately for all cases. There is no reason to suspect that all cases would turn out as favorably as the results in Table 5. About the strongest conclusion that can be reached with the presently analyzed data is that the results favor the hypothesis of negative correlation between estimated confidence range and self-rating.

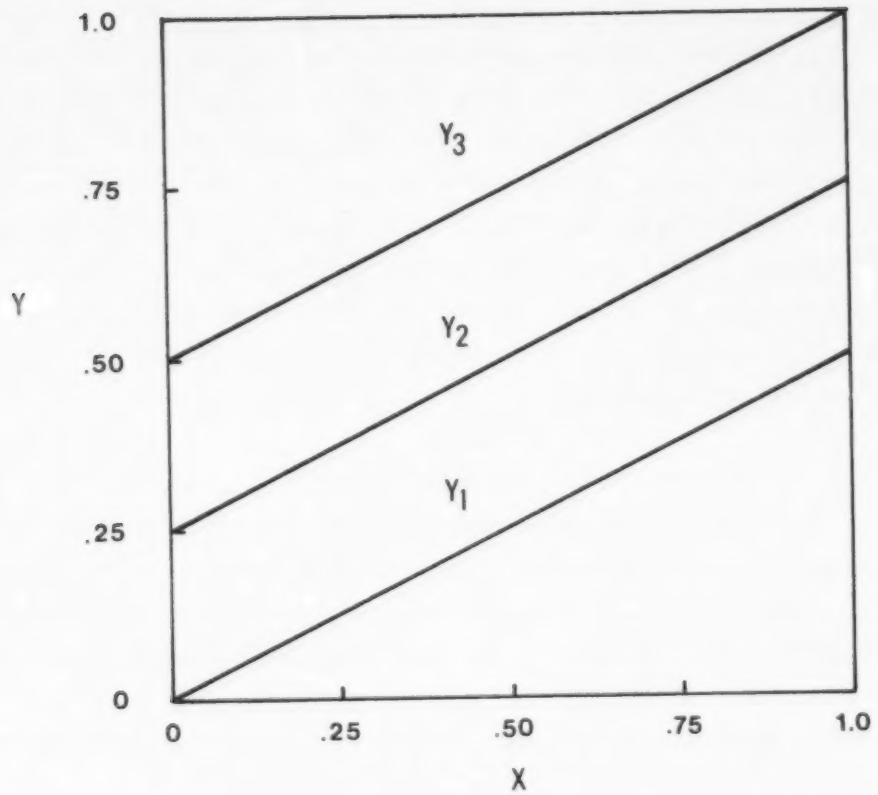


FIG. 18 ILLUSTRATION OF REDUCTION IN CORRELATION WITH POOLED POPULATIONS

TABLE 5: Correlation Between R and Δy, for NO_2, Disability, Children

Subpopulation	$r_{R\Delta y}$
0%	-.507
10%	-.445
50%	-.612
90%	-.626

Correlations were also run for the three group indices, $\Delta\bar{y} = \bar{Y}_u - \bar{Y}_l$ (difference between the mean of the upper limit and mean of the lower limit), s, and $\bar{R}$ (average self-rating). Cases for these correlations were pooled across percent population and population type, giving 56 data points per correlation. This produced 9 different correlations for each pair type. These are displayed in Table 6.

As might be expected with aggregated data, correlations in Table 6 are generally higher than those in Table 4. This contrary to the effect of pooling data which, because of the virtual constancy of R across populations, decreases correlation; aggregating data by taking averages tends to increase the correlations. However, a slightly higher percentage have a different sign from the expected; self-rating should be negatively correlated with both estimated confidence range and standard deviation; estimated confidence range should be positively correlated with standard deviation. Thirteen of the twenty-seven correlations are significant at the .05 level (one-tailed test); five of these have the wrong sign.

TABLE 6: Correlations Between $(\Delta\bar{y},s)$, $(\Delta\bar{y},\bar{R})$, $(s,\bar{R})$

Pollutant	Impairment	$(\Delta\bar{y},s)$	$(\Delta\bar{y},\bar{R})$	$(s,\bar{R})$
OX	I	.197	.025	.284*
	Da	.158	.016	.214
	Dc	.162	.069	-.237*
NO_2	I	-.247*	-.142	.317*
	Da	-.274*	-.022	-.309*
	Dc	-.274*	.109	-.366*
CO	I	.369*	-.280*	-.142
	Da	.534*	-.242*	-.259*
	Dc	.215	-.216	-.010

* Significant at .05 level (one-tailed test)

I = Incapacity

Da = Disability

Dc - Discomfort

Figure 19 graphs $\Delta\bar{y}$ against $\bar{s}$; Figure 20 graphs $\Delta\bar{y}$ against $\bar{m}$. From Figure 20 it is evident that the panelists are scaling their confidence ranges roughly in proportion to their X estimates. The correlation between Median X and $\Delta\bar{y}$ (Median Y_u - Median Y_l) for all Oxidant cases except 0% population (126 cases) is .61. There is thus a discrepancy between the logarithmic scaling of the concentration estimates, and the arithmetic scaling of the confidence ranges. In addition, from Figure 20 it can be observed that the slope of the relationship between m and Log (Y_u - Y_l) is different (lower) within pollutants than it is across pollutants. This explains, in part, the peculiar within-pollutant behavior of $\Delta\bar{y}$ in Figure 19. Since s also scales on the log transform, it appears likely that the negative correlations between $\Delta\bar{y}$ and s for the NO_2 cases in Table 6 arise from the same discrepancy between log scaling on s and arithmetic scaling on Y_u - Y_l.

Table 7 is a display of average R broken out in percent population and degree of impairment categories. There is an evident increase of self-ratings between estimates for 10% population and 90% population. This is an interesting trend which might be paraphrased by the statement that the respondents link their assurance in their estimates with the percentage of the population they expect to be affected. One hypothesis might be that this is a form of synaesthesia -- percent population is one scale for "certainty". Otherwise Table 7 shows no clear pattern that can be related to Δy or s. These is one feature on which all three indices agree, namely that the estimates for NO_2 are less certain than those for OX and CO. The overall averages for the three indices for each of the three pollutants are shown in Table 8. The comparison of OX and CO is a virtual standoff, essential equality on R, greater uncertainty for OX on Δy, and greater uncertainty for CO on s.

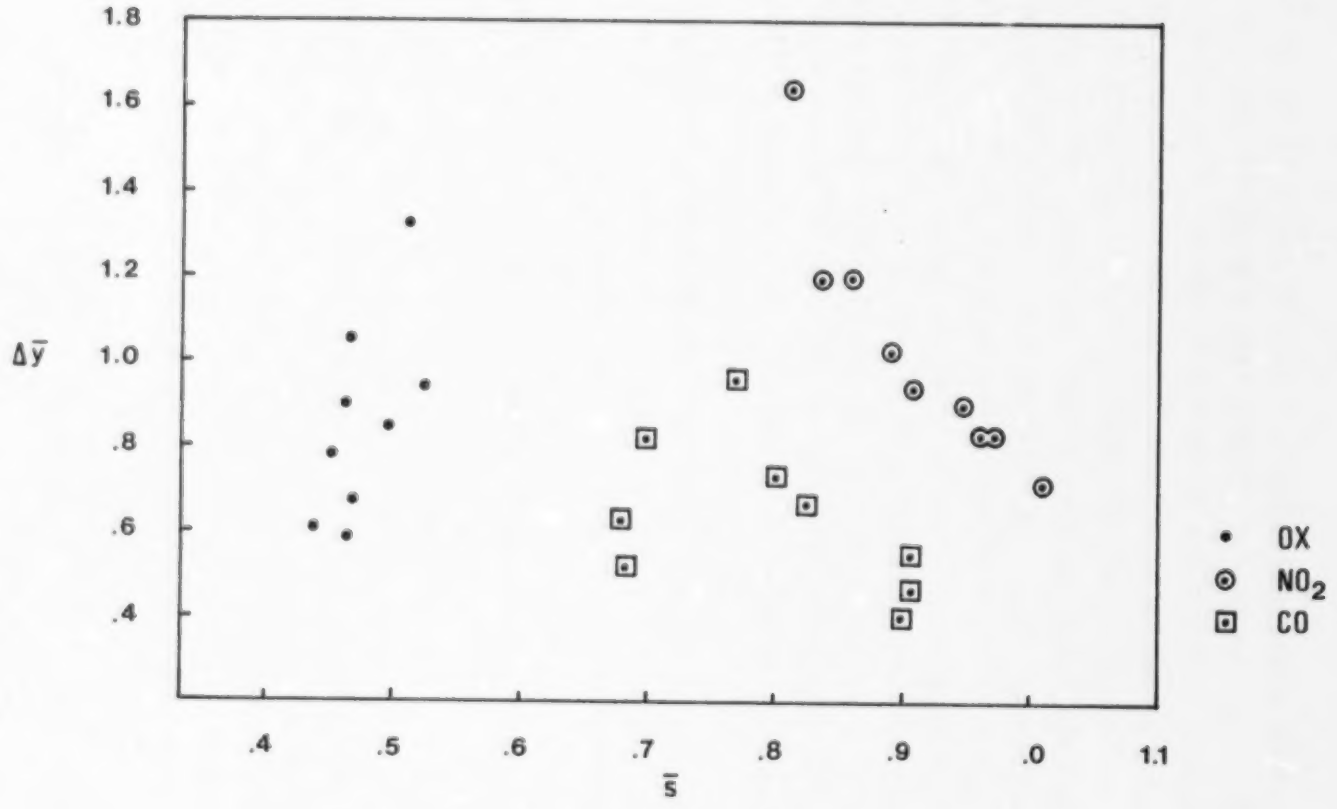


FIG. 19 AVERAGE ESTIMATED INTERVAL vs OBSERVED STANDARD DEVIATION

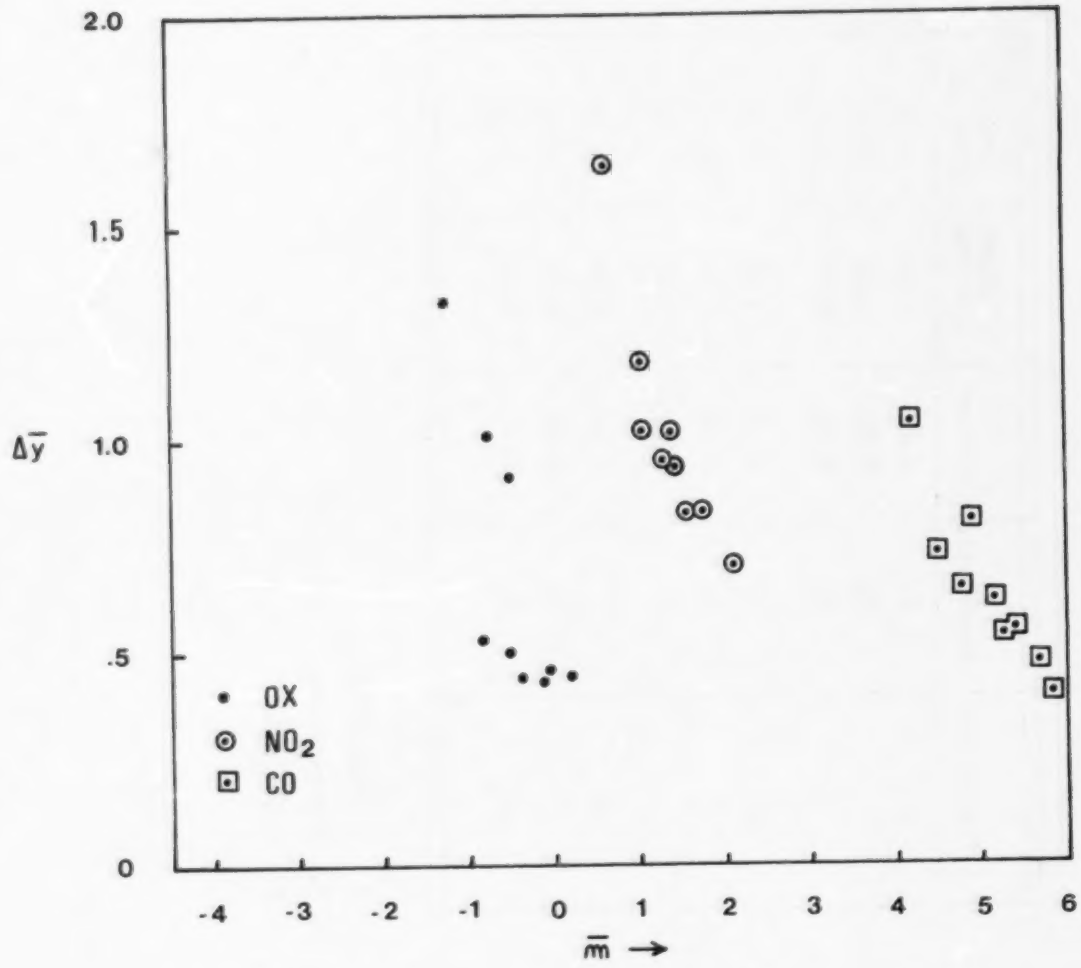


FIG. 20 AVERAGE ESTIMATED CONFIDENCE INTERVAL vs AVERAGE LOG MEAN

TABLE 7: Average $\bar{R}$ by Percent Population, Degree of Impairment and Pollutant Type.

% Population	Impairment	Oxidant	Nitrogen Dioxide	Carbon Monoxide
10	I	4.07	3.91	4.36
	Da	4.48	3.84	4.21
	Dc	4.65	4.44	4.26
50	I	4.26	3.97	4.79
	Da	4.47	3.90	4.61
	Dc	4.64	4.16	4.59
90	I	4.44	4.23	4.96
	Da	4.64	4.01	4.76
	Dc	5.17	4.14	4.73

I = Incapacity

Da = Disability

Dc = Discomfort

TABLE 8: Overall Averages for Δy, R and s

	OX	NO_2	CO
R	4.54	4.07	4.59
Δy	.86	1.03	.64
s	.47	.91	.80

This is a difficult subsection to summarize. The theory of errors gives a central role to the standard deviation, and to this extent, it might be considered the most reliable of the three indices. Despite some anomalous behavior within pollutants, it exhibits a fairly clear scaling property with m (Figure 13) which Δy does not (Figure 20). If we assume that s is roughly the average of the indivudual standard deviations then Δy should be about 3.28s. For a normal distribution, the 95^{th} percentile occurs at $m + 1.64\sigma$, and the 5^{th} percentile at $m - 1.64\sigma$. Thus the 90% confidence range is 3.28σ. Since Y_u is defined as the dosage level which the respondent thought no more than 5% of cases would exceed, with a corresponding definition for Y_l, $Y_u - Y_l$ should correspond to the 90% confidence range. The ratio $\Delta y/3.28s$ for the overall averages of these two indices is displayed in Table 9.

TABLE 9: $\Delta y/3.28s$ For Three Pollutants

OX	NO_2	CO
.52	.35	.24

These ratios are well in line with Capen's conclusion that estimated confidence ranges are underestimates by a factor of two or more[13].

E. An Estimation Model. Inspection of the panel responses shows, as one might expect, systematic variations from case to case. Simple logic dictates that estimates for the various population percentages should increase with the percentage. Similarly, the estimates should increase with the degree of impairment; for the same percentage population, $x(I) > x(Da) > x(Dc)$. A somewhat more subtle issue is, whether estimates for the various populations consist simply of scaled versions of each other, or whether each population is a special case.

There are two points of view from which models of the estimated data can be approached. One is the standard view that the panel members are trying to approximate an implicit and informal model of the underlying phenomena. The other is to view the data as expressing a model of the estimation process, that is, as resulting from psychological scaling operations that are loosely tied to the actual phenomena. Without extensive objective data for comparison, there is no decisive way to distinguish between these two possibilities. Observed regularities in the data may result from common perceptions of the physiological reactions to inhaled pollutants, or they may result from common modes of judgment.

Delphi practitioners in the past have not addressed this issue. By and large, estimates for complex, interrelated quantities have been treated as independent (or at least as separate) judgments, and the data analyzed estimate by estimate. This is clearly unsatisfactory where a sequence of estimates are closely related, as in the HHD study. There is no "canonical" treatment for such data.

The present investigation is in a sense a pioneering effort in this area. As it turns out, it appears useful to borrow elements from both points of view to impose a meaningful structure on the data.

In the HHD study, one elementary model was proposed to rationalize the percent population estimates, namely, the threshold model for the onset of a given complex of symptoms. The model assumes that for a specified set of symptoms, and a specified pollutant, each member of the relevant population can be characterized by a critical value (threshold) of the dosage, beyond which he will exhibit the symptoms. The threshold has a distribution within the population which determines the fraction of the population that will exhibit the symptoms at a given dosage. On a priori grounds it was suggested that the distribution was likely to be lognormal. This led to the model

$$\frac{dF}{dc} = \frac{\phi(z)}{c} \qquad z = A + B \log c \tag{14}$$

where F is the fraction of the population affected, c is the dosage (concentration experienced for one hour). $\phi(z)$ is the normal density function, and A and B are constants which depend on the type of pollutant, population, and degree of impairment. This is often called a probit model.

For computational convenience, (14) was approximated by the logistic model

$$\log \frac{F}{1-F} = C + D \log c \tag{15}$$

Since both the probit model and the logit model have natural zeros at 0, and since the panel estimates for 0% population were somewhat erratic and exhibited large standard deviations, only the 10%, 50% and 90% estimates were used to fit the model (least squares fit to the constants). As might be expected, it was not difficult to find relatively good fits in most cases to these three points, and no persuasive confirmation of the model was attempted. It was treated more as a convenient method of analyzing the data and extrapolating the given estimates to other population percentages.

Actually, both the logit and probit models have an elementary consequence that can be used to formulate a useful test of fit. The consequence is

$$\log c(50\%) = 1/2(\log c(10\%) + \log c(90\%)) \tag{16}$$

For the probit model, this follows from the symmetry of $\Phi(z)$; i.e., $z(90\%) = -z(10\%)$, $z(50\%) = 0$. The same symmetry holds for the logit model, since $\log\frac{F}{1-F} = -\log\frac{1-F}{F}$ and $\log .5/.5 = \log 1 = 0$. The difference between the left hand side of (16) and the right hand side is thus a good measure of error for either model. Its application to the HHD data will be discussed below.

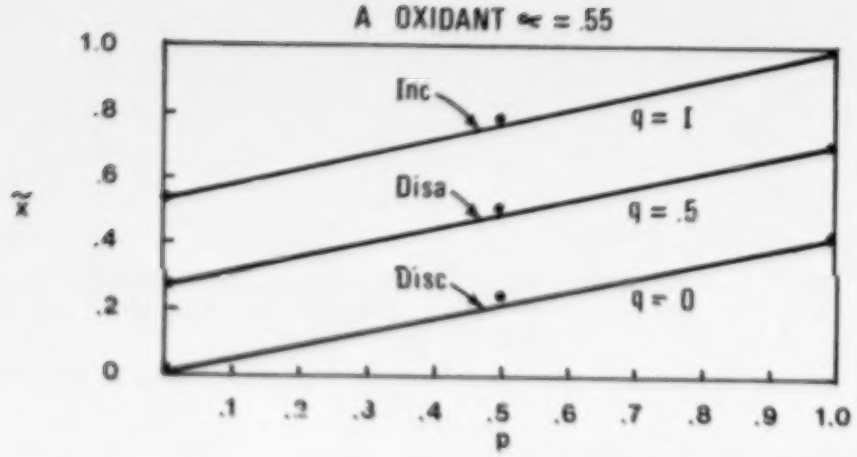
It is apparent on inspection of the tabulated data, that regularities exist among the estimates for different levels of impairment. However, the symptoms defining the levels of impairment were specified separately for each pollutant and each population type. Thus, there is no obvious physical scale to which the levels of impairment can be attached. In fact, it is not obvious that it is physically meaningful to refer to such a scale. On the other hand, the terms "Incapacity", "Disability", and "Discomfort" were employed as common descriptors across populations. As it turns out, the respondents appear to have interpreted these three labels as defining three points on a scale which is very similar to the percent population scale; i.e., log c(Da) is roughly halfway between log c(I) and log c(Dc). In this instance, a relatively stable and consistent psychological scale appears to have determined the estimates for different degrees of impairment.

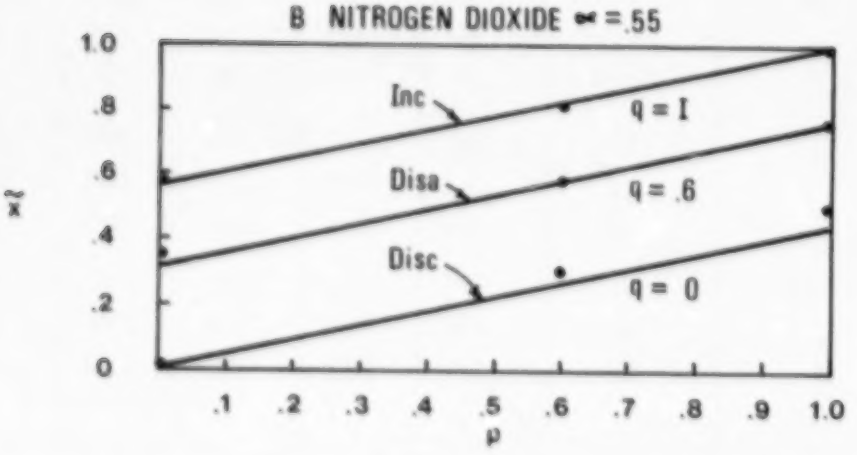
The values of $\tilde{x} = \frac{x - x(Dc,10\%)}{x(I,90\%) - x(Dc,10\%)}$, averaged over populations for each pollutant, are displayed in Table 10. The standard deviation of $\tilde{x}$ is $\tilde{s}$ computed across populations. As can be seen, the $\tilde{s}$ are generally small, indicating that the average values tabulated are a good approximation to the separate population cases.

The same data is displayed in graphical forms in Figures 21 (A)-(C), with $\tilde{x}$ plotted against a rescaling of the percent population labelled p. Since $\tilde{x}$ has been normalized by dividing by $x(I,90\%) - x(Dc,10\%)$, $p(10\%) = 0$ and $p(90\%) = 1$. From the discussion of Equation 16 above, one would suspect that a best fit would be obtained by setting $p(50\%) = .5$. However, a slightly better fit is obtained by setting $p(50\%) = .6$ for NO_2 and CO. To this extent, the condition prescribed by Equation 16 is not strictly met.

TABLE 10: Normalized Estimates and Standard Deviation by Percent Population, Degree of Impairment, and Pollutant Type.

% Population	Impairment	OX		NO_2		CO	
		$\tilde{x}$	$\tilde{s}$	$\tilde{x}$	$\tilde{s}$	$\tilde{x}$	$\tilde{s}$
10	I	.589	.040	.563	.033	.767	.028
	Da	.343	.050	.282	.052	.427	.004
	Dc	0	0	0	0	0	0
50	I	.807	.040	.789	.008	.909	.016
	Da	.581	.070	.516	.047	.580	.022
	Dc	.296	.038	.253	.036	.176	.017
90	I	1.0	0	1.0	0	1.0	0
	Da	.762	.065	.715	.046	.667	.009
	Dc	.491	.038	.431	.035	.314	.024





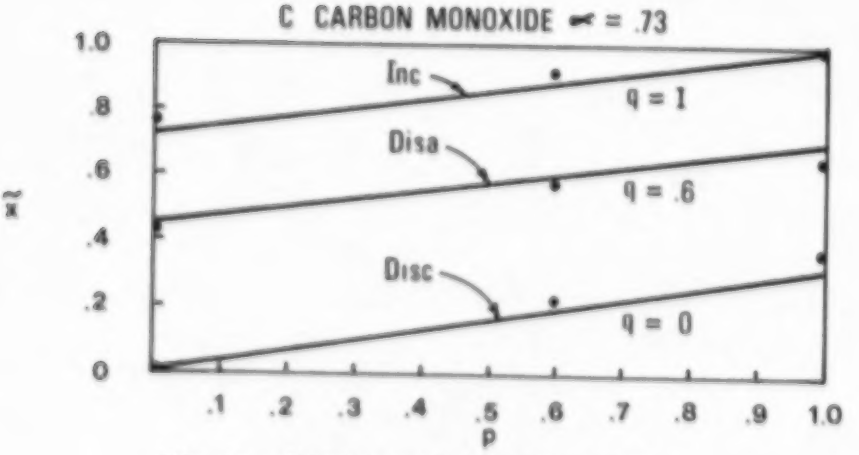


FIG. 21 NORMALIZED ESTIMATE $\tilde{x}$ AS A FUNCTION OF THE PARAMETER p

The straight lines in Figures 21 (A) to (C) are obtained by the relationship

$$\tilde{x} = \alpha p + (1 - \alpha)q \qquad (17)$$

Where q is a scaling on degree of impairment; q(I) = 1, q(Dc) = 0, and q(Da) = .6 for NO_2 and CO. The analogy with the percent population scale is quite striking. α is a parameter which expresses the relative weight given to percent population and to degree of impairment in estimating x. The normalized estimates $\tilde{x}$ for different pollutants expressed as a function of the parameter are shown in Figures 21 (A) to (C), where $\alpha(OX) = \alpha(NO_2) = .55$ and $\alpha(CO) = .73$.

No attempt was made to optimize any of the parameters in the construction of Figure 21. Simplicity of both form and content have been the major criteria. In this spirit, Figure 21 (A) was constructed setting p(50%) = q(Da) = .5. A surprisingly good fit to the OX data is obtained with this simple scaling.

Since there are only three general cases (the three popputants) for establishing the model, the question whether the simpler scaling p(50%) = q(Da) = α = .5 would fit other pollutants to an acceptable degree of approximation remains a viable option. It seems probable that for many purposes the simple model would be adequate.

Perhaps the model, as developed to this point, would be more easily understood described in terms of the method of implementation. For each pollutant, and each population type, panelists would be asked to estimate three numbers: X(Dc,10%), X(I,90%) and α. It will probably be clear that α is a constant for each pollutant; in which case, it need not be estimated for every population type. From these three numbers, the constants A and B in Equation 14 can be determined, and the X for any other degree of impairment, percent population combination can be calculated. Auxilliary estimates, such as self-ratings or confidence ranges, would be elicited only for the three estimated numbers. Notice that the model replaces the nine X estimates (for a given pollutant and population type) in the HHD study with two, or at most three, numbers.

An interesting area for further exploration is the possible extension of the psychological scale of level of impairment to a continuum, rather than the three discrete levels specified in the present study. The data indicate that an interval scale, with reference points specified by defining Discomfort and Incapacity in terms of particular symptoms, should be relatively easy to construct. Some modification of the present model might be needed if the extension included conditions outside the Discomfort-Incapacity interval. One useful application of this scale would be to generate cumulative distributions of degrees of impairment in a given population for a designated dosage. Figure 22 (a) and (b) are speculative illustrations of this application, based loosely on the present data within the I-Dc interval. The ordinate expresses the percent of the given population exhibiting at least the degree of impairment indicated on the abscissa, for the specified dosage. For example, in Figure 22 (b), for c = 1.9, the graph indicates that 70% of the population will exhibit Disability or worse.

One additional question relating to mcdelling was investigated. There is the possibility that a rough scale of "severity of illness" for the various subpopulations determined to a large extent the required dosage estimates. This exploration ran into a thorny problem. The differences between estimates across populations are in most cases small compared with differences within populations (e.g., between different degrees of impairment). As a result, small errors can seriously influence across-population scaling. Errors can be expected from the variability of panelists' judgments; but in addition the published "raw" data ccntains a number of recording errors arising either from transcription or from carelessness or the part of respondents. Some of these can be identified by noting, e.g., that the X estimate is outside the Y_u, Y_l range, or that X(50%) is less than X(0%) or that X(Normal) is less than X(I11), etc. A pleasant irony of the development of the model described above is that in examining some egregious discrepancies with the model, a number of errors of this sort were uncovered.

To ameliorate the problem of variability with small differences, and hopefully to average out some of the recording errors, the x estimates (group

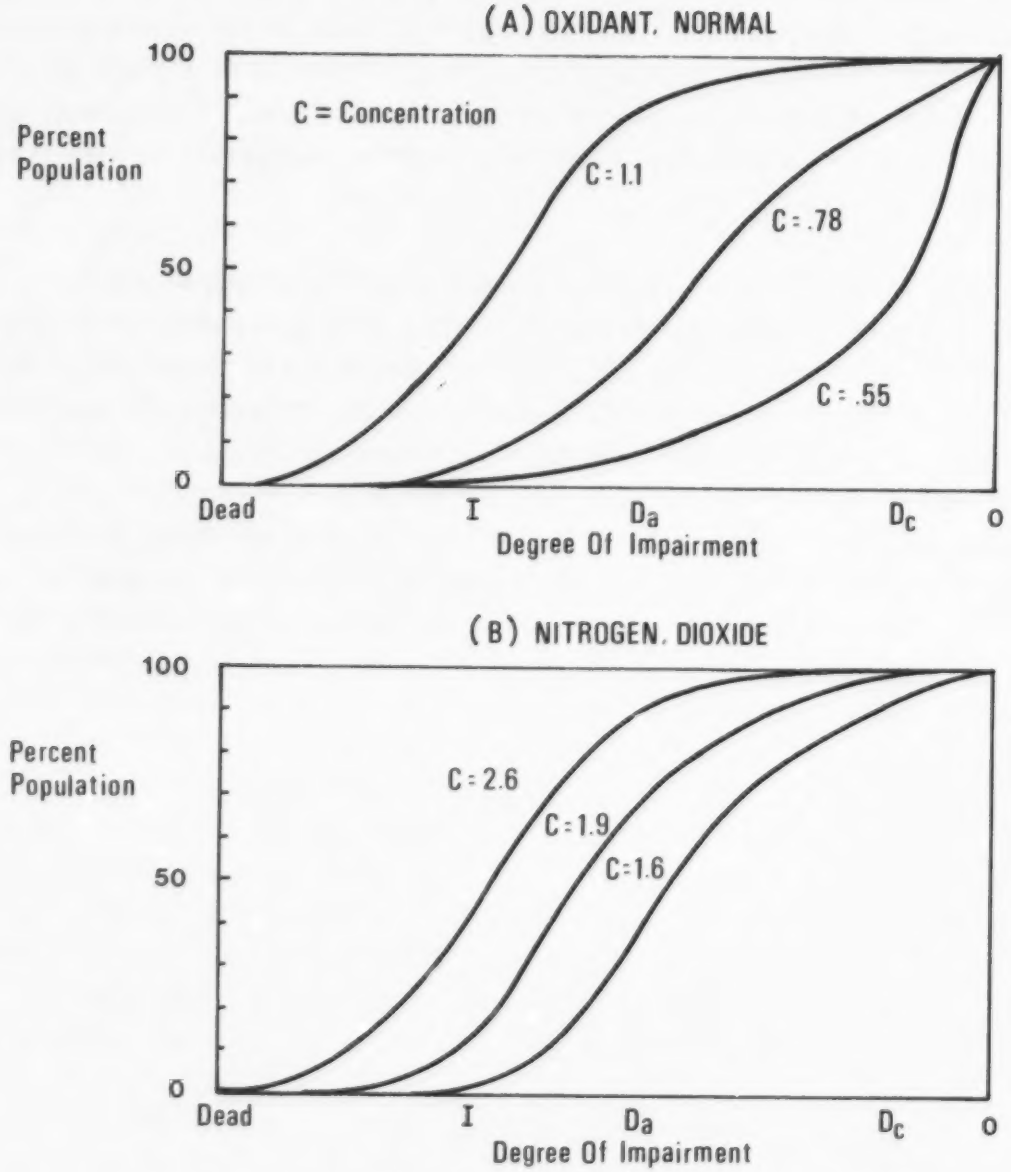


FIG. 22 ILLUSTRATION OF CUMULATIVE DISTRIBUTION

averages) were averaged across all percent population, degree of impairment cases except 0 percent population for each population type. This generated what could be labelled $\bar{\bar{x}}$. $\bar{\bar{x}}$ is thus a kind of index of the degree of severity (of illness) associated with each population. Figure 23 is a graph of $\bar{\bar{x}}$ (Normal) - $\bar{\bar{x}}$ for the 14 populations and the three pollutant types. The connecting lines are not intended to show functional relationships, but merely to keep track of the three pollutants.

It is clear from Figure 23 that there is a fair amount of similarity in the way in which "severity" is judged for the three pollutants, after some special cases are discounted. These are primarily the two heart conditions, 4 and 5, for CO, and Hay Fever and Asthma, 6 and 9, for OX. Otherwise the qualitative behavior of the index is similar for the different pollutants. Kendall's coefficient of concordance, W, across the populations is .70, $\chi^2 = 27.3$, $p<.01$. However, the numerical relative severity does not appear to be sufficiently stable to justify trying to introduce population type into the model on the basis of the present data. Figure 23 does justify some optimism concerning the possibility of formulating a general severity scale that could account for a large part of the variance over populations.

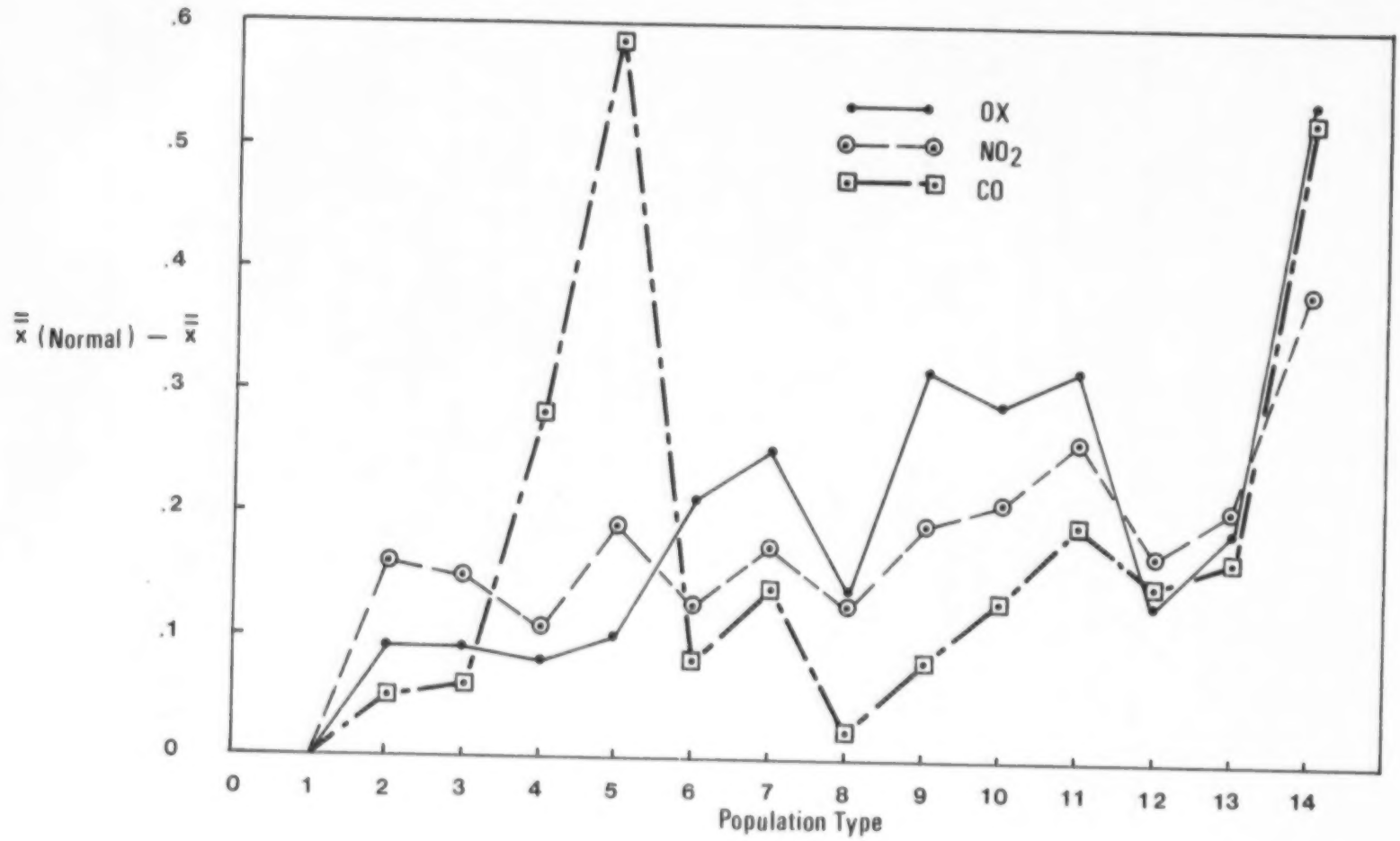


FIG. 23 VARIATION OF $\bar{x}$ WITH POPULATION

SECTION 8

DISCUSSION

Despite the large number of estimates elicited during the HHD study, the high degree of dependence among the estimates reduces the amount of statistical information by a large but difficult to specify factor. It probably would be desirable in future studies of this sort to introduce estimates of two sorts not included in HHD, namely, (1) cases (e.g., other types of pollutants) where it is likely beforehand that the panel members are relatively poorly informed and (2) cases (if available) where the panel members probably can make very good estimates. The purpose of these additional estimates, of course, is not to obtain information about the subject matter, but rather to "calibrate" standard deviations, self-ratings and estimated confidence ranges. If at all possible, the additional, calibration estimates should cover a relatively wide range of size of estimates.

The overriding issue, if the HHD study is to be considered relevant to practical decisions, is whether the analogy between the HHD data and the data from the Rand studies is sufficiently close so that the large bias observed in the latter can be imputed to the HHD estimates. Since the analogy looks reasonably close, the hypothesis of lognormality of distributions appears plausible, and the scaling of standard deviation with mean is compatible with the psychonumeric hypothesis. The prudent conclusion is probably the cautious one, namely that the HHD estimates contain about the same proportion of bias as the Rand estimates. This entails the assumption that the bias is about twice the expected error computed from the observed standard deviation.

How this assumption is to be translated into an operational assessment of a given estimate depends in part on the role accorded to the psychonumeric hypothesis in defining error. If respondents scale their estimates on the logarithm, then both standard deviation and expected error will increase

exponentially with the size of the estimate. This effect will not be invidious if the relevant phenomena fit the same sort of scaling law. As an example, the psychologically harmful effects of noise increase as the logarithm of the physical intensity. Presumably, the relevant errors are those in the psychological scale, not in the physical scale. Similarly, the relevant physiological effects of air pollutants may increase as the logarithm of the concentration. The model derived from the HHD estimates appears to be saying just this. Both the susceptability (threshold) and the degree of impairment are proportional to the logarithm of the dosage. The relevant error would appear to be in terms of these effects, not in terms of the concentration.

On the other hand, it is clear that at the present time policy is defined in terms of the physical scales (e.g., in attaching various kinds of alerts to various concentration levels). If decisions are focussed on concentration levels, then a difficult problem of "correcting for" the psychonumeric phenomenon is posed. Rescaling is not a serious problem; what is serious is correcting for the exponential increase in expected error with size of the estimate.

The issue posed by these considerations clearly goes well beyond the application of the theory of errors to group estimates, and involves both the substantive nature of the effects of air pollution, and relevant policy variables. Neither of these are within the competence of this report. If the scale of primary interest is determined to be the concentration, then, rather than the wholesale rejection of estimates that would result from imposing the criterion $s \leq .5$ (As pointed out above, none of the NO_2 or CO estimates meet this criterion.), a more sagacious procedure would be to publish the entire set of estimates with their attendent indices (s, $\overline{R}$, $\overline{\Delta y}$) and a brief explanation of their significance.

The picture concerning the usefulness of the estimated confidence range and the self-rating does not emerge sharply in the HHD data. In part this results from the small variation of the average self-rating across populations. The correlations between Δy, R, and s weakly support the hypothesis that they

are measuring something in common, but this is highly obscured by anomalous and at present unexplained variations within related sets of estimates. Without some explanation for these anomalies, it is probably unreasonable to base a decision concerning the trustworthiness of a given estimate on such variations.

One rather firm conclusion would appear to be that the estimated confidence range cannot be used to specify absolute confidence limits (e.g., confidence limits intended to guide policy). However, the results of the analysis do not rule out the possibility that they have some utility in measuring relative degree of certainty. That is, it seems unlikely that 90% of all actual cases, if they could be established by experiment or field trial, would lie between the estimated Y_u and Y_l limits. However, for two different estimates for the same pollutant, but, e.g., for different populations, if the panel estimated a larger ΔY for one estimate than for the other, then on the average the estimate with the larger ΔY can be expected to be less reliable than the other. This statement is expressed in terms of the untransformed estimates, rather than in terms of the log transform Δy, because of the linear scaling on the confidence estimates discussed in Section 7,D.

Perhaps the most positive outcome of the present analysis is the identification of a relatively simple model which appears to account for most of the variation dependent on the percentage population and degree of impairment variables. Although a number of interesting questions are raised by the model -- e.g., whether some distribution other than the lognormal or the logistic would give a better fit -- these appear to be issues of fine-tuning. Considering the large amount of randomness in the estimates that one would expect from the theory of errors, given the size of the standard deviations, the model does suprisingly well. The model opens the possibility of formulating a scale of degree of impairment which would cover a much wider range of symptom states than those embodied in the terms Incapacity, Disability, and Discomfort, and from what can be gathered from the present data, generalization of the model to fit this more comprehensive scale should not be a large step.

REFERENCES AND NOTES

1. Leung, S. K.,E. Goldstein, N. Dalkey, Draft Final Report: Human Health Damages from Mobile Source Air Pollution . EPA Contract No. 68-01-1809, by California Air Resources Board, March 1975.

2. The theory of uncertain estimates as probabilistic judgments has been elaborated under the terms Bayesian estimates, subjective probability, personal probability. Relevant names are F. P. Ramsey, B. De Finneti, L. J. Savage, W. Edwards, A good exposition of the approach can be found in L. J. Savage, Foundations of Statistics, John Wiley & Sons, 1954. The theory of estimates as a model has been employed by many cognitive psychologists, including E. Brunswick, P. Hoffman, K. Hammond, L. Goldberg, Robyn Dawes. A convenient reference is P. Hoffman, "Cue-Consistency and Configurality in Human Judgment", in Formal Representation of Human Judgment, B. Kleinmuntz, ed. John Wiley and Sons, 1968.

3. Aitchison, J.and J. A. C. Brown, The Lognormal Distribution, Cambridge Univ. Press, 1957.

4. Aitchison, J. and J. A. C. Brown, ibid., Chap. 5.

5. Dalkey, N.and Bernice Brown, Comparison of Group Judgment Techniques with Short-Range Prediction and Almanac Questions. The Rand Corporation, R-678-ARPA, May 1971.

6. Dalkey, N. An Experimental Study of Group Opinion . Futures, Sept., 1969, pp 408-426.

7. Dalkey, N. Group Decision Analysis. To be published, winter 1976.

8. Stevens, Vide, S. S. Ratio Scales of Opinion. In D. K. Whitla, ed., Handbook of Measurement and Assessment in Behavioral Science, Addison-Wesley, 1968.

9. Raimi, R. A. The Peculair Distribution of First Digits. Scientific American, Dec., 1969, pp 109-120.

10. A number of relevant studies on this topic concerning weather forecasting and almanac type judgments were presented at the Conference on Bayesian Research, Los Angeles, Cal., 1976. Especially pertinent was the report, "Do Those Who Know More Also Know More About How Much They Know?", by Sarah Lichtenstein and Baruch Fischhoff of the Oregon Research Institute.

11. This topic has received extensive investigation. The issues were first presented in an unpublished paper by H. Raiffa and M. Alpert, Harvard, 1967, and have been followed up by R. Winkler, J. J. Selvidge, T. Brown, and others. Of particular interest is a recent study by E. C. Capen, "The Difficulty of Assessing Uncertainty", presented at the 50th Annual Fall Meeting of the Society of Petroleum Engineers of AIME, Dallas, Texas, Sept. 28 - Oct. 1, 1975.

12. Capen, E. C. The Difficulty of Assessing Uncertainty. Presented at the 50th Annual Fall Meeting of the Society of Petroleum Engineers of AIME, Dallas, Texas, Sept. 28 - Oct. 1, 1975.

13. Capen, E. C., R. V. Clalp, Wm. M. Campbell Competitive Bidding in High-Risk Situations. Jour. Petroleum Technology, June 1971, pp. 641, 653.